Serviços de Computação em Nuvem

As melhores estratégias para uma infraestrutura mais eficiente, performática e acessível

Casa do Código | alura

CLAYTON EDUARDO DOS SANTOS

Sumário

- ISBN
- Sobre o livro
- Sobre o autor
- Parte 1: Contextualização teórica
  - 1 Introdução
  - 2 Computação em nuvem
- Parte 2: Primeiros passos
  - 3 Precificação
  - 4 Serviços de autenticação, autorização e auditoria
- Parte 3: Serviços básicos
  - 5 Serviços de computação em nuvem
  - 6 Serviços de armazenamento
  - 7 Serviços de rede
- Parte 4: Serviços gerenciados pelo provedor de nuvem
  - 8 Serviços de monitoração
  - 9 Serviços gerenciados pelo provedor de nuvem
- Parte 5: Segurança e laboratórios práticos
  - 10 Segurança e conformidade
  - 11 Discussões e conclusões

ISBN

Impresso: 978-85-5519-400-9

Digital: 978-85-5519-399-6

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Santos, Clayton Eduardo dos
Serviços de computação em nuvem : as melhores estratégias para um infraestrutura mais eficiente, performática e acessível / Clayton Eduardo dos Santos. -- 1. ed. -- São Paulo : AOVS Sistemas de Informática, 2025.

Bibliografia.
ISBN 978-85-5519-400-9

1. Ciência da computação 2. Computação em nuvem
3. Computação em nuvem - Medidas de segurança
4. Infraestrutura de tecnologia da informação
I. Título.

25-278368 CDD-004

Índices para catálogo sistemático:

1. Ciência da computação 004

Aline Graziele Benitez - Bibliotecária - CRB-1/3129

Caso você deseje submeter alguma errata ou sugestão, acesse http://erratas.casadocodigo.com.br.

Sobre o livro

A inclusão digital da população desempenhou papel fundamental na consolidação da digitalização dos negócios, ao passo que tecnologias como a computação em nuvem tornaram possível o surgimento de empresas que utilizam a transformação digital como principal diferencial. Tal estratégia permite ao empreendedor aproveitar eventuais janelas de oportunidade, uma vez que a tecnologia em questão viabiliza maior agilidade na implantação do negócio. No entanto, a utilização de serviços baseados em nuvem por um período de tempo mais longo exige um entendimento mais amplo do paradigma por parte dos profissionais envolvidos. Nesse tipo de cenário, as escolhas realizadas no processo de implantação devem ser revistas periodicamente, com um olhar voltado à eficiência e à inovação, sem que se percam de vista o valor agregado e a manutenção da vantagem competitiva. Para tanto, podem ser utilizadas diferentes estratégias, que vão desde a mudança do modelo de contratação de serviços, passando pela adoção de soluções *multicloud*, ou ainda a eventual utilização de um modelo de nuvem híbrida, por exemplo.

Como o livro está organizado

Os conteúdos abordados na presente obra foram organizados em uma sequência lógica que busca, sobretudo, a didática. Inicialmente, você terá acesso a conceitos introdutórios importantes, que vão se tornando mais específicos e detalhados à medida que a leitura avança. O texto está dividido em onze capítulos, a saber:

- **Capítulo 1 - Introdução:** Traz uma visão geral sobre o tema, com uma dinâmica que prepara o leitor ou leitora para os capítulos posteriores;
- **Capítulo 2 - Computação em nuvem:** Apresenta o conceito de computação em nuvem e suas principais características;
- **Capítulo 3 - Precificação:** Ilustra a importância do uso de ferramentas de precificação disponibilizadas pelos provedores

de nuvem, bem como de estratégias para tornar a adoção da tecnologia economicamente viável;
- **Capítulo 4 - Serviços de autenticação, autorização e auditoria:** Introduz o conceito de políticas de acesso e como aplicá-las aos usuários, grupos e demais entidades relacionadas;
- **Capítulo 5 - Serviços de computação em nuvem:** Define e apresenta os principais tipos de serviços de computação ofertados em soluções de nuvem privada e pelos provedores de nuvem pública;
- **Capítulo 6 - Serviços de armazenamento:** De forma análoga ao capítulo anterior, são apresentados os principais tipos de serviços relacionados ao armazenamento de dados;
- **Capítulo 7 - Serviços de rede:** Encerrando o ciclo dos "serviços básicos de nuvem", são apresentados os serviços relacionados à gestão de redes nesse tipo de cenário;
- **Capítulo 8 - Serviços de monitoração:** Discute a importância dos serviços de monitoração em ambientes de nuvem;
- **Capítulo 9 - Serviços gerenciados pelo provedor de nuvem:** Sugere alternativas à gestão tradicional de serviços, deixando-a a cargo do provedor;
- **Capítulo 10 - Segurança e conformidade:** Correlaciona soluções de segurança amplamente utilizadas no modelo de infraestrutura própria com as soluções equivalentes ofertadas no modelo de nuvem;
- **Capítulo 11 - Discussões e conclusões:** São apresentadas tendências, perspectivas futuras e conclusões relacionadas aos tópicos discutidos no decorrer da obra.

Para quem este livro é indicado

Ao contrário do que se possa imaginar, o entendimento do conceito, funcionamento e utilização de tecnologias de computação em nuvem não é mais uma exclusividade restrita a profissionais atuantes na área de desenvolvimento, infraestrutura e afins. Os reflexos promovidos pela transformação digital exigem que colaboradores de

diferentes setores conheçam o paradigma, entendam suas particularidades e, em especial, tirem proveito das oportunidades disponibilizadas pelo modelo.

Ciente dessa realidade, a presente obra foi redigida com o objetivo de contemplar toda e qualquer pessoa que tenha a necessidade de entender como realizar a transição do modelo de infraestrutura própria para o modelo de infraestrutura em nuvem, considerando os principais tipos de implantação e categorias de serviços disponíveis, com um olhar que leva em consideração o ponto de vista do usuário, elencando as principais dúvidas, decisões e experimentações típicas desse tipo de cenário. Ao final de cada capítulo, você tem acesso a exercícios e materiais complementares, disponíveis *online*, que auxiliam na fixação e no entendimento dos conteúdos.

Sobre o autor

Professor Doutor Clayton Eduardo dos Santos

Figura 0.1: Professor Clayton Eduardo dos Santos

Formação acadêmica

Possui mestrado e doutorado em Engenharia Elétrica, obtidos na Escola de Engenharia de São Carlos da Universidade de São Paulo (USP), especialização em Gestão Pública pela Universidade Católica Dom Bosco, graduação em Análise e Desenvolvimento de Sistemas pelo Centro Universitário Claretiano e licenciatura em Matemática pelo Centro Universitário de Rio Preto, respectivamente.

Experiência profissional

Desde 2008, é professor efetivo no campus Bragança Paulista do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, na área de Redes de Computadores. É instrutor credenciado nas academias AWS Academy, Cisco Networking Academy, Huawei ICT Academy, Oracle Academy, Red Hat Academy e VMware IT Academy.

Contatos

Plataforma	URL
Currículo Lattes	http://lattes.cnpq.br/6246109453810593
LinkedIn	https://www.linkedin.com/in/claytones/

Agradecimentos

O processo de construção de um livro vai muito além da simples redação dos capítulos.

Em primeiro lugar, é preciso encontrar um tema sobre o qual se deseja falar, dentre tantos assuntos possíveis. A escolha do tema de computação em nuvem foi motivada pelo meu envolvimento cotidiano com diversos programas acadêmicos promovidos por grandes *players* da área de tecnologia, aos quais o campus Bragança Paulista do IFSP está vinculado. Agradeço a todos os gestores desses programas, com quem tive a oportunidade de trabalhar, pela seriedade, pelas oportunidades e, obviamente, pelo conhecimento adquirido. Um agradecimento especial ao professor e amigo Cristiano Alves Pessoa, responsável por trazer a primeira parceria do tipo para a instituição e, sobretudo, por me convidar a participar do processo de implantação.

Vencida a primeira etapa, inicia-se o processo de se "colocar no papel" as ideias que, em um primeiro momento, parecem muito claras e objetivas, mas que, durante o processo, por vezes, se mostram turvas e distantes. Agradeço ao apoio da minha esposa e parceira de todas as horas, Ana Paula Müller Giancoli, não só por entender a motivação e a dedicação necessárias para o desenvolvimento de um trabalho realizado por meses a fio, em horários vagos cada vez mais alternativos e raros, como também por fazer parte dele, construindo com carinho, criatividade e entusiasmo cada uma das imagens que tanto enriqueceram os textos.

Por fim, é importante ter um público-alvo em vista. Meus atuais e futuros alunos, presenciais e remotos, foram, sem dúvida, meu maior incentivo para iniciar esta jornada. Ter um material de autoria própria, a ser utilizado como referência em diferentes cursos, sempre foi algo que me chamou a atenção. Agradeço a todos que,

em uma ou mais oportunidades, estiveram comigo em sala, laboratório ou em aulas síncronas ministradas em diferentes plataformas online.

Meus agradecimentos também à Casa do Código, pela oportunidade de escrever meu primeiro livro e pelo suporte durante toda a jornada. Vivian Matsui, sua mentoria e profissionalismo foram fundamentais para a conclusão deste trabalho. Espero que possamos trabalhar juntos novamente, em livros que tratem de novos temas.

Dedico este trabalho a meus pais, Antônio dos Santos (*in memoriam*) e Maria Aparecida Valarine dos Santos, a quem agradeço pela dedicação, pelos exemplos e, especialmente, pela referência que representam em minha vida.

Muito obrigado!

Parte 1: Contextualização teórica

Capítulo 1
Introdução

A rotina de um profissional da área de tecnologia da informação exige atualização constante. A rigor, os recursos computacionais existem, sobretudo, para resolver problemas e melhorar processos, considerando diferentes níveis de complexidade e cargas de trabalho. Aliado a essa questão, temos de levar em conta o dinamismo da área. Novas tecnologias surgem todos os dias, prometendo resolver os problemas do cotidiano de maneira mais eficiente do que as ferramentas já existentes. Embora isso, em muitos casos, seja verdade, a transição por parte dos usuários nem sempre é simples.

Com o advento da transformação digital, o investimento em tecnologia da informação deixou para trás a alcunha de "custo", passando a ocupar papel de destaque na disputada prateleira dos chamados "investimentos estratégicos". Tal protagonismo traz grandes responsabilidades, pois, administrativamente, grandes investimentos obviamente pressupõem lucros ainda maiores.

Frente a esse cenário, o(a) profissional da área lida em seu cotidiano com uma linha tênue que separa a inovação e o bom senso. Para explicar melhor essa afirmação, vamos imaginar que uma empresa recentemente tenha realizado um aporte financeiro significativo para a construção de um *datacenter* próprio. Posteriormente, a popularidade da tecnologia de computação em nuvem e a crescente oferta de provisionamento de infraestrutura como serviço por parte dos provedores de nuvem pública certamente chamariam a atenção dos gestores, e o aporte seria objeto de questionamentos ao setor

que sugeriu o investimento na construção de infraestrutura própria, *on-premises*. Cabe ao profissional, inicialmente, esclarecer que a inovação é inevitável e faz parte do dinamismo da área, além de sugerir estratégias para aproveitar melhor os investimentos já feitos, sem deixar de lado a inovação tecnológica.

Nesse caso hipotético, em especial, a infraestrutura preexistente poderia tirar proveito dos benefícios da tecnologia de computação em nuvem ao adotar uma solução de nuvem privada, por exemplo. Com o passar do tempo, recursos que seriam destinados à atualização da infraestrutura local poderiam naturalmente ser destinados ao provisionamento de novos serviços em ambiente de nuvem pública, caracterizando um ambiente de nuvem híbrida que utiliza o melhor dos dois mundos.

Dessa forma, a eventual transição de uma infraestrutura própria para a nuvem pública pode e deve ocorrer no momento certo, com análise constante dos fatores que influenciam essa decisão — especialmente o custo total de propriedade e as melhores estratégias para aquisição e manutenção na nuvem —, visando à eficiência e à economia, e evitando movimentos de repatriação de dados motivados por erros estratégicos.

Para que produtos e serviços oferecidos pela indústria possam ser disponibilizados para consumo utilizando a grande vitrine digital proporcionada pela internet, é importante que, de tempos em tempos, a indústria encontre maneiras de tornar a utilização dos recursos computacionais mais eficiente, performática e acessível. A computação em nuvem surge com esse propósito e oferece aos usuários a possibilidade de desfrutar de serviços básicos de armazenamento, computação e transferência de dados no formato de serviços, ou, ainda, preservando investimentos legados que, por algum motivo, precisam ser mantidos na infraestrutura das empresas.

Partindo do princípio de que a tecnologia em questão tende a se tornar cada dia mais presente, a atualização dos profissionais de

diferentes segmentos que vão interagir com esse tipo de infraestrutura em suas atividades cotidianas é um fator que deve ser levado em consideração. Nesse processo de capacitação, é inevitável encontrarmos diferentes guias de estudos voltados à preparação para exames de certificação profissional disponíveis no mercado. Tais certificações possuem diferentes níveis de aprofundamento e são ofertadas pela grande maioria das empresas de tecnologia, que, por sua vez, são responsáveis pelos principais provedores de nuvem pública disponíveis atualmente. As oportunidades disponíveis no mercado tendem a variar, podendo priorizar contratações de especialistas em determinado *vendor*, de acordo com os requisitos de cada projeto.

Considerando o cenário supracitado, é de fundamental importância que a pessoa candidata tenha em mente que a tecnologia utilizada pelos provedores de nuvem é, essencialmente, a mesma. Todos os chamados "serviços básicos", divididos em diferentes categorias, estão presentes em todos eles, inclusive em produtos que viabilizam a implementação de nuvens privadas. O que muda é o nome de cada serviço, que tende a ser comercializado como um produto que valorize a marca à qual está atrelado.

Para o(a) profissional que deseja se atualizar, essa é uma excelente notícia, pois, uma vez que se tenha entendido os conceitos e funcionalidades de cada uma das ferramentas pertencentes às principais categorias de serviço disponíveis, automaticamente a pessoa estudante da tecnologia estará apta a administrar soluções equivalentes, disponibilizadas por outros provedores de nuvem, com uma curta curva de aprendizado.

Principais provedores de nuvem pública

Atualmente, clientes interessados em provisionar sua infraestrutura "além dos muros da organização" podem contar com provedores de nuvem pública consolidados no mercado. Dentre os principais, destacam-se:

Provedor	*Website*
Alibaba Cloud	https://www.alibabacloud.com/
Amazon Web Services	https://aws.amazon.com/pt/
Google Cloud Platform	https://cloud.google.com/
Huawei Cloud	https://www.huaweicloud.com/intl/pt-br/
IBM Cloud	https://www.ibm.com/br-pt/cloud
Microsoft Azure	https://azure.microsoft.com/pt-br/
Oracle Cloud Infrastructure	https://www.oracle.com/br/cloud/

Evidentemente, o portfólio de produtos ofertado pelos provedores de serviços deve variar de acordo com o tempo de funcionamento destes, bem como em função das áreas em que possuam especialidade notória. Apesar das eventuais particularidades, é importante lembrar que todos eles, invariavelmente, disponibilizam "serviços básicos" equivalentes nas categorias de computação, armazenamento e redes. Da mesma maneira, serviços de apoio essenciais, como os relacionados aos processos de autenticação, autorização, auditoria, monitoração e precificação, dentre outros, também fazem parte do catálogo de serviços disponibilizado por cada um.

Principais soluções de nuvem privada

Existem organizações que, por diferentes razões, optam pela adoção do modelo de infraestrutura própria, de forma completa, ou parcial. Nesse tipo de cenário, recomenda-se a utilização de componentes de *software* que, utilizando a infraestrutura própria legada, preexistente, implementam uma solução privada *on-premises* capaz de disponibilizar aos usuários um portfólio de serviços básicos

equivalente aos tipicamente encontrados em infraestruturas provisionadas em ambientes de nuvem pública.

Atualmente, as principais soluções voltadas à implementação de nuvens privadas são:

Produto	*Website*
Apache CloudStack	https://cloudstack.apache.org/
OpenStack	https://www.openstack.org/

1.1 Considerações importantes

Como vamos trabalhar

Conforme discutido anteriormente, provedores de nuvem pública e soluções de *software* voltadas à implementação de nuvens privadas disponibilizam um mesmo conjunto de recursos básicos essenciais, com funcionalidades equivalentes. Dentro de cada categoria de serviço disponível no portfólio dos provedores, existem ferramentas específicas com funções especializadas, voltadas ao atendimento de diferentes demandas, regras de negócios e fluxos de trabalho. Dentre os chamados "serviços básicos" supracitados, destacam-se:

Categoria de Serviço	Funções
Serviços de Armazenamento	Armazenamento de arquivos Armazenamento em blocos Armazenamento de objetos

Categoria de Serviço	Funções
Serviços de Computação	Baseados em contêineres Baseados em computação sem servidor Baseados em virtualização
Serviços de Rede	ACLs Grupos de segurança Balanceadores de carga Subredes VPCs

Nos próximos capítulos, vamos conhecer de forma mais detalhada as principais características das categorias de serviço consideradas indispensáveis em ambientes de nuvem, bem como seus principais componentes. Para que você compreenda os conceitos envolvidos de maneira clara e objetiva, serão inicialmente apresentados os referenciais técnicos referentes a cada um dos serviços abordados, bem como os principais procedimentos, parâmetros de configuração e práticas necessárias para uma implementação real. Para ilustrar esse processo, utilizaremos as ferramentas disponibilizadas pela *Huawei Cloud*, com vistas à apresentação do console de gerenciamento *web*, interface que permite a interação com a infraestrutura a ser provisionada em ambiente de nuvem.

Vale ressaltar que os exemplos utilizados ao longo dos capítulos podem ser facilmente adaptados para outros provedores de nuvem ou soluções de nuvem privada, uma vez que os conceitos apresentados tratam o tema de forma ampla, tornando possível o entendimento necessário para aplicação prática dos conhecimentos adquiridos em ambientes *multicloud*.

Conteúdo adicional

Ao final de cada capítulo, são disponibilizados questionários *online*, com perguntas pertinentes ao tema apresentado, com vistas à fixação dos conteúdos mais importantes. A opção pelo conteúdo externo é motivada pela utilização de um ambiente similar aos encontrados em exames de certificação oficiais, bem como pela viabilidade de atualização das questões de acordo com mudanças eventualmente ocorridas na tecnologia. Da mesma maneira, a seção de recursos didáticos complementa o conteúdo do capítulo com sugestões de leituras baseadas em artigos, documentações oficiais e atualizações técnicas e conceituais inerentes ao dinamismo da área.

O objetivo da presente obra é apresentar uma visão geral sobre o tema computação em nuvem. A abordagem utilizada foi idealizada de modo a permitir que o leitor ou leitora iniciante sinta-se confortável com o tema e, em especial, torne-se apto(a) a utilizar a tecnologia imediatamente após a leitura do material.

Para complementar os conteúdos apresentados, caso tenha interesse, você terá acesso gratuito a um curso de extensão ministrado pelo autor e ofertado pelo campus Bragança Paulista do Instituto Federal de São Paulo, na modalidade remota, em parceria com o programa *Huawei ICT Academy*. O treinamento em questão é preparatório para a certificação profissional de serviços de nuvem, nível *associate*, ou simplesmente *"HCIA - Cloud Services"*, oferecida pelo fabricante, com certificado de conclusão emitido pela plataforma *Huawei Talent*. As informações necessárias para acesso ao treinamento estão disponíveis na seção de recursos didáticos do último capítulo.

Bons estudos!

Capítulo 2
Computação em nuvem

Em um passado relativamente recente, "digitalizar" um pequeno ou médio negócio, em tese, significava sistematizar a gestão de processos, utilizando recursos informáticos para tanto. Em geral, a empresa, ainda que de forma analógica, já existia, estava consolidada, ou em vias de se estabelecer. O processo de informatização trazia benefícios significativos relacionados à agilidade, diminuição de erros humanos, maior controle sobre as operações realizadas, entre outras contribuições. Tendo em vista o pequeno número de equipamentos necessários nesse tipo de cenário, os investimentos realizados normalmente apresentavam rápido retorno, uma vez que os *softwares*, computadores, equipamentos de comunicação e demais periféricos normalmente eram utilizados em âmbito local, no próprio estabelecimento.

Naturalmente, tais benefícios passaram a ser diferenciais perceptíveis na implantação de novos negócios e na modernização dos já existentes. Com a internet, as empresas passaram a adotar arquiteturas mais robustas, baseadas no modelo matriz e filial, que trouxeram maior complexidade e, consequentemente, a necessidade de maiores investimentos para o provisionamento da infraestrutura computacional.

A inclusão digital da população, aliada aos atrativos benefícios da transformação digital, fizeram da internet a principal vitrine de produtos e serviços do planeta, conquistando públicos de diversos locais, faixas etárias e classes sociais.

Frente a esse novo cenário, atualmente até mesmo os micro e pequenos negócios já nascem digitais. Como isso é possível?

2.1 Despesas de consumo e despesas de capital

Ao lidar com o provisionamento de infraestrutura, é preciso ter pleno entendimento de que a implantação de um *datacenter* próprio implica em custos que vão muito além dos já esperados investimentos em equipamentos de computação, armazenamento e conectividade. O chamado "custo total de propriedade", ou simplesmente TCO (*Total Cost of Ownership*), contempla outras despesas significativas, a saber:

- Custos imobiliários;
- Energia elétrica;
- Equipe de TI especializada;
- Mobiliário;
- Redundância de serviços (energia, equipamentos de armazenamento, computação e conectividade, *links*, sistema de refrigeração e afins);
- Refrigeração;
- Segurança patrimonial;
- Segurança perimetral.

Fazendo uma breve reflexão, consumimos alimentos e não precisamos plantar ou cuidar de animais em nossas casas, bebemos água potável e não temos uma estação de tratamento em nossos quintais. Da mesma maneira, dispomos de energia elétrica sem nos preocuparmos em gerir uma usina, assim como navegamos na internet sem a necessidade de construir um provedor de acesso, além de telefonarmos sem a preocupação de administrar uma empresa de telecomunicações. Então, por qual motivo precisamos construir e administrar um *datacenter* próprio para que possamos ter acesso a serviços de computação, armazenamento e transferência de dados, com qualidade?

A resposta para essa pergunta é simples: não precisamos! Enfim, vamos entender as principais motivações da tão falada computação em nuvem.

Para entender melhor esse cenário, cabe aqui a definição de dois termos importantes que vamos utilizar nos próximos capítulos, *CapEx* e *OpEx*. O primeiro é uma abreviação da expressão *Capital Expenditures*, ao passo que o segundo refere-se a *Operational Expenditures*. As chamadas despesas de capital referem-se a investimentos de grande volume, normalmente relacionados à aquisição de bens permanentes, como imóveis, equipamentos e demais demandas resultantes da composição de infraestrutura. Já as despesas operacionais estão relacionadas a bens de consumo, ou seja, despesas operacionais realizadas em nosso cotidiano.

Uma vez colocadas à mesa as regras do jogo, basta decidir qual tipo de modelo gostaríamos de adotar para promover a transformação digital do nosso novo empreendimento. O primeiro, no qual precisaremos de tempo e de um montante de recursos financeiros significativo, suficientes para a construção e gestão do nosso próprio *datacenter*; ou o segundo, no qual teremos acesso imediato a todos os recursos e funcionalidades desejadas, devidamente disponibilizadas por um provedor de computação em nuvem, que ao final do mês tarifará os serviços utilizados de acordo com o perfil de consumo do usuário, tal como acontece nas contas mensais que recebemos em nossas casas ou empresas, enviadas aos clientes pelos fornecedores dos serviços de água, luz ou telefone.

E então, qual é a melhor escolha?

2.2 A definição de computação em nuvem, segundo o NIST

Peter Mell e Timothy Grance, do *National Institute of Standards and Technology - NIST*, definiram a computação em nuvem como "*um modelo que permite acesso global, conveniente e sob demanda via rede, a um conjunto compartilhado de recursos de computação configuráveis (como redes, servidores, armazenamento, aplicativos e serviços, por exemplo) que podem ser rapidamente provisionados e liberados com o mínimo esforço de gerenciamento ou interação com o provedor de serviços*". (Fonte: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf)

Certamente, existem outras definições que descrevem o termo de forma competente, no entanto, a referência supracitada é amplamente adotada na literatura, o que torna natural sua utilização também por aqui. Então vamos analisar os principais termos utilizados pelos autores, de modo a facilitar o entendimento dos conceitos envolvidos e contextualizá-los com nosso cotidiano.

Definir a computação em nuvem como um "modelo" é algo especialmente interessante. Estamos falando de uma abordagem que pode e deve ser reproduzida, tanto por empresas que desejam explorar esse tipo de oportunidade comercialmente, como por aquelas que eventualmente pretendam dar uma sobrevida a seus ativos computacionais legados, compatibilizando sua infraestrutura própria com o novo paradigma e tornando-a mais moderna, proveitosa e expansível.

Permitir "acesso global", via rede, aos recursos provisionados na nuvem é um pré-requisito factível via diferentes tecnologias de conectividade, de banda larga móvel ou fixa, disponíveis nos dias atuais. Serviços especializados de segurança são aplicados em diferentes camadas, de forma complementar e de modo a garantir a confidencialidade, integridade e disponibilidade das informações trafegadas.

O termo "conveniente" pode ser associado às diferentes interfaces de acesso utilizadas pelo usuário para interação com os recursos

provisionados remotamente, no ambiente de nuvem. De acordo com os requisitos de projeto, existem alternativas que podem atender desde a usuários que necessitam de uma abordagem interativa tradicional, baseada na utilização de um console de gerenciamento *web*, passando por soluções mais específicas, voltadas a eventuais necessidades de automação, implementadas com a utilização de scripts executados via interfaces de linha de comando, até abordagens mais dinâmicas que consomem os recursos programaticamente, via aplicativos *web* ou móveis.

A utilização de recursos "sob demanda" é um dos fatores determinantes para o sucesso da computação em nuvem. Nesse tipo de modalidade de oferta de serviços, o usuário consome "*um conjunto compartilhado de recursos de computação*", conforme sua necessidade e efetiva utilização, sem precisar aderir a contratos ou planos preestabelecidos pelo fornecedor, assim como faz com outros serviços cotidianos essenciais, como água, energia elétrica ou telefone. A arquitetura utilizada é baseada em *web services*, utilizando o mecanismo de requisições e respostas, o que torna o consumo de serviços mensurável, transparente e auditável.

A gerência de "recursos configuráveis" confere ao usuário da tecnologia o poder de adequar sua infraestrutura de acordo com suas necessidades ou em função do aumento ou diminuição da demanda real, a qualquer tempo, de forma rápida e à distância de um clique.

Tecnologias de virtualização de infraestrutura e redes, bem como ambientes de armazenamento de dados dedicados e especializados, possibilitam que os recursos sejam "*rapidamente provisionados e liberados com o mínimo esforço de gerenciamento*" da forma mais conveniente ao usuário, utilizando um modelo de autoatendimento sob demanda, conforme descrito anteriormente.

Em termos práticos, os chamados "provedores de serviços" administram uma grande quantidade de *datacenters* físicos distribuídos globalmente, munidos de recursos virtualmente

inesgotáveis, de modo a atender a um público crescente, com demandas e conformidades de negócio específicas.

Tipos de implantação de infraestrutura em nuvem em função da localização dos recursos

De forma geral, o paradigma de computação em nuvem disponibiliza a seus usuários a oportunidade de consumir recursos computacionais essenciais como serviço. Para tanto, existem diferentes tipos de implantação, que variam de acordo com o propósito desejado, conforme a seguir.

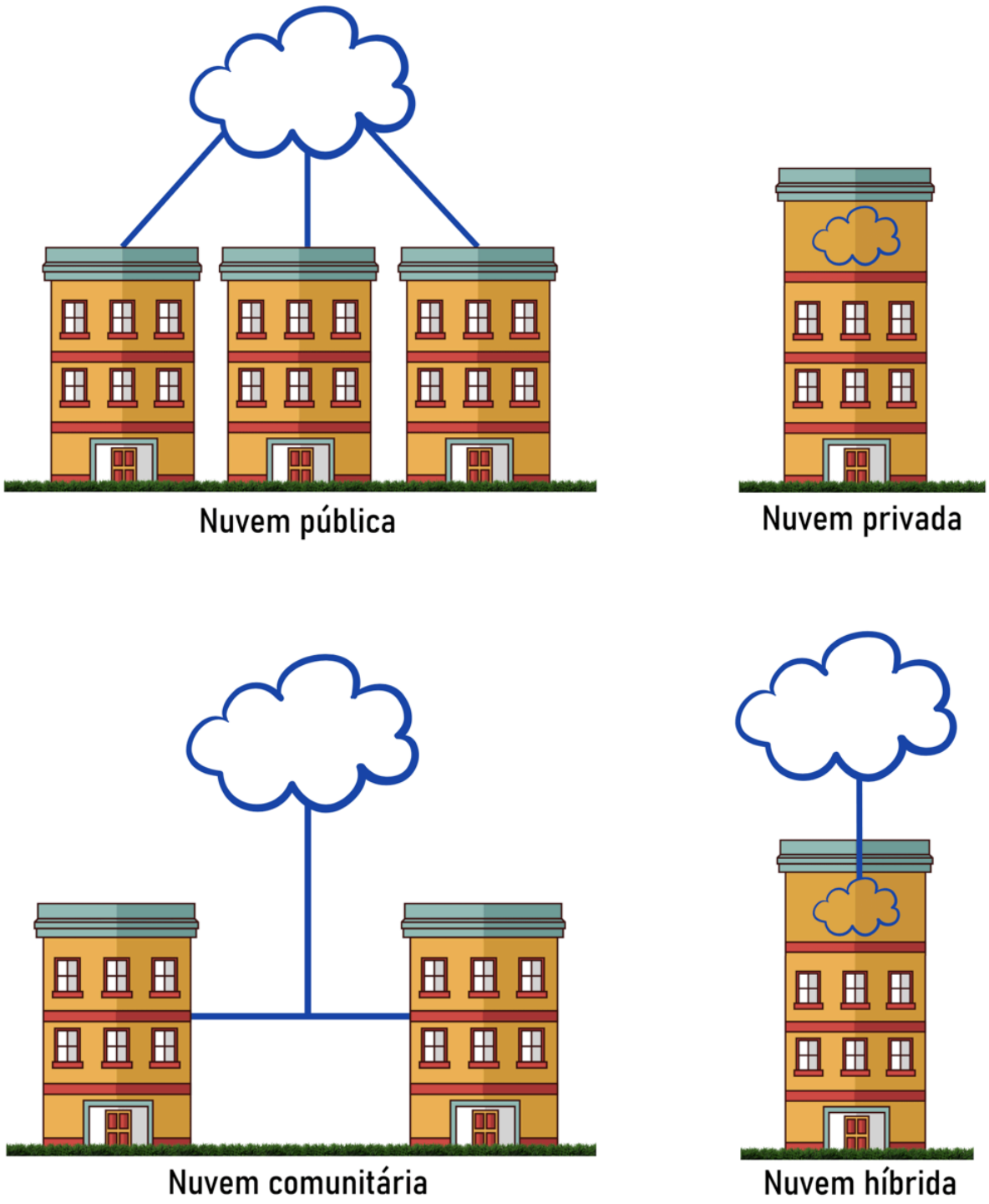


Figura 2.1: Tipos de implantação

- **Nuvem pública:** nesse tipo de implantação, pretende-se atingir um público-alvo amplo e diversificado, tendo em vista a natureza comercial, acadêmica ou governamental, que

normalmente caracteriza as instituições que optam pela adoção do referido modelo. Aqui, por definição, toda a infraestrutura fica alocada nas instalações do provedor de nuvem;

- **Nuvem privada:** como o próprio nome sugere, o modelo de nuvem privada é de propriedade, responsabilidade e uso exclusivo de uma única organização. Em geral, trata-se de um tipo de implantação que atende a um perfil de público que busca dar sobrevida a uma infraestrutura própria legada, para preservação de investimentos realizados previamente em equipamentos e licenças de *software*, por questões de conformidade ou, ainda, com a intenção de usufruir de recursos intrínsecos à tecnologia localmente, como escalabilidade, alta disponibilidade e recuperação de desastres;

- **Nuvem comunitária:** busca atender a um conjunto de entidades, instituições ou comunidades com interesses e objetivos comuns. Pode pertencer, ser gerida e operada por uma das organizações pertencentes à comunidade, ou ainda ser composta por uma federação de nuvens privadas, pertencentes a cada uma das participantes;

- **Nuvem híbrida:** por se tratar de um modelo comum, intercambiável e independente de fornecedor, a computação em nuvem permite a combinação de dois ou mais tipos de implantação (pública, privada ou comunitária), proporcionando ao cliente e aos usuários usufruírem das vantagens de cada uma delas.

Modelos de serviço utilizados na computação em nuvem

Ao optar pela utilização da computação em nuvem, é importante que o usuário da tecnologia esteja ciente de sua responsabilidade na gestão de sua nova infraestrutura.

Conforme dito anteriormente, em uma eventual escolha pela implantação de uma nuvem privada, todas as responsabilidades ficam a cargo da organização.

Por outro lado, ao decidir-se pela utilização de um provedor de nuvem pública, o usuário passa a contar com o chamado *Modelo de responsabilidade compartilhada*, no qual as atividades de gerência são divididas entre o provedor de serviços e o cliente, podendo variar de acordo com o modelo de serviço adotado.

Figura 2.2: Modelo de responsabilidade compartilhada

A computação em nuvem utiliza três principais modelos de serviço, a saber:

- **Infraestrutura como Serviço (*IaaS*):** nessa modalidade, o usuário utiliza os serviços básicos de computação, armazenamento e transferência de dados de sua preferência, sem se preocupar com a gestão da infraestrutura física adjacente, que é de responsabilidade do provedor de serviços. No entanto, a gestão de todas as camadas superiores, a partir do sistema operacional, é de responsabilidade do cliente. Dentre as tarefas a serem realizadas, podemos citar: criação de usuários, instalação e configuração de *softwares*, bibliotecas e serviços em geral, instalação de *patches* de segurança, definição e implantação de políticas de segurança, entre outras atribuições;
- **Plataforma como Serviço (*PaaS*):** aqui, o usuário faz uso de serviços que são totalmente gerenciados pelo provedor. A responsabilidade do cliente limita-se ao fornecimento do código-fonte da aplicação a ser provisionada, bem como das informações que devem ser populadas no sistema de gerenciamento de banco de dados escolhido. Todas as operações necessárias para que os serviços supracitados sejam provisionados, permanecendo seguros e disponíveis, são de responsabilidade do provedor de serviços. É importante que o usuário interessado nessa modalidade de serviço atente-se às linguagens de programação e aos *engines* de bancos de dados suportados pelo provedor de serviços;
- **Software como Serviço (*SaaS*):** aplicativos com gestão centralizada, hospedados e disponibilizados pelo provedor de nuvem e parceiros comerciais são consumidos pelo usuário que tem como única responsabilidade a gestão de suas credenciais de acesso, em geral, usuário e senha.

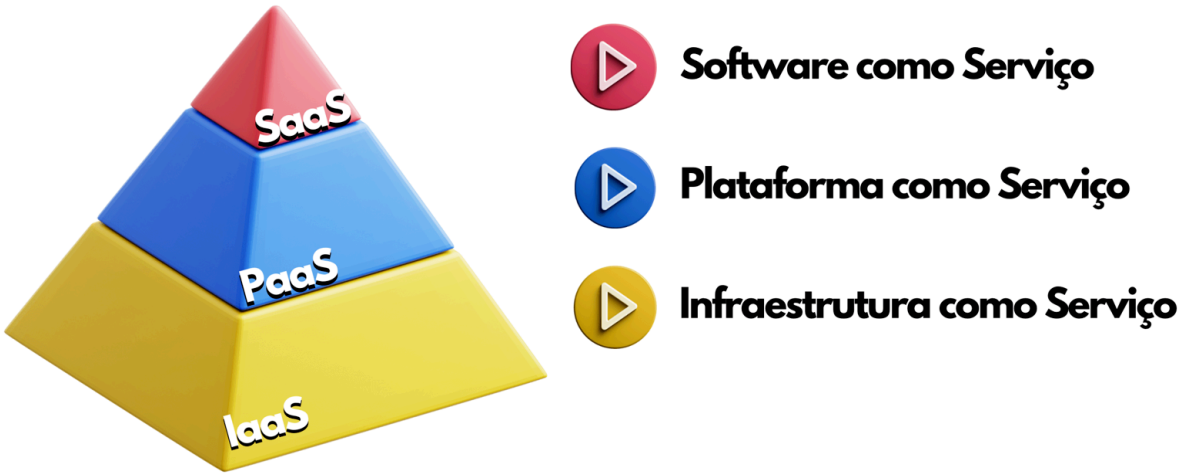


Figura 2.3: Modelos de serviço

Vantagens da computação em nuvem

Dentre as vantagens da utilização da computação em nuvem como alternativa ao modelo de infraestrutura própria, destacam-se:

- **Alcance:** uma empresa que, em um primeiro momento, optou por concentrar sua infraestrutura em uma única região pode, de forma rápida e a qualquer tempo, replicar seus ativos computacionais para outras regiões, dentro do mesmo país, continente ou qualquer outra localidade atendida pelo provedor de nuvem, com vistas à experiência do usuário. Alternativamente, podem ser utilizados serviços que realizam o *cache* de informações estáticas utilizadas com maior frequência, possibilitando resultados semelhantes com melhor custo-benefício, uma vez que a infraestrutura original permanece a mesma;

- **Agilidade:** negócios de ocasião, em geral, exigem respostas rápidas nos processos de tomada de decisão e provisionamento de recursos, o que dificilmente é factível no modelo de infraestrutura própria. A utilização da computação em nuvem possibilita aos empreendedores o aproveitamento de eventuais janelas de oportunidade, tendo em vista que os recursos

necessários são provisionados em questão de minutos, via console ou programaticamente;
- **Economia:** trocar despesas fixas de capital por despesas de consumo variáveis faculta às empresas a opção de investirem o montante economizado em seus produtos e clientes, aumentando a satisfação dos colaboradores e do público-alvo. Da mesma maneira, tecnologias mais modernas e efetivas são oferecidas sem custo adicional ao cliente em substituição a outras consideradas obsoletas, tendo em vista a economia de escala à qual o provedor de nuvem tem acesso. Um exemplo clássico para esse caso de uso é a substituição de discos mecânicos tradicionais, popularmente conhecidos como HDs, pelos discos de estado sólido, ou SSDs, nos serviços de armazenamento disponibilizados pelos provedores de nuvem;

- **Elasticidade:** no modelo de infraestrutura própria, administradores de sistemas precisam estabelecer, de forma teórica, os picos de utilização máxima de recursos, para que possam adquirir seus equipamentos de computação, armazenamento e redes. No modelo de nuvem, é possível condicionar a variação do "tamanho" da infraestrutura em função da demanda de recursos real de forma dinâmica. Tal recurso, denominado elasticidade, possibilita eventuais ajustes de dimensionamento dos ativos computacionais em qualquer momento, de forma automática, quando associados a determinado evento ou programada, em datas e horas específicas que historicamente apresentam maior demanda de acessos.

Infraestrutura global dos provedores de nuvem pública

Os provedores de nuvem pública possuem uma ampla estrutura de *datacenters* distribuídos globalmente. Por convenção, a grande maioria destes utiliza o termo "região" para caracterizar sua presença física e consequente atuação comercial em determinado país. Em geral, cada região é interligada a duas outras regiões

vizinhas, que se conectam com mais duas regiões diferentes cada, até que a malha de conectividade esteja completa utilizando sempre essa mesma lógica.

Dessa maneira, a redundância de conectividade é garantida, eliminando os pontos únicos de falha de cada segmento, em caso de queda de um dos *links* de comunicação, por exemplo. O provedor passa a contar com uma infraestrutura global, capaz de estabelecer transferência de dados entre as diferentes regiões devidamente atendidas, sem utilizar trânsito de dados via internet pública, tendo em vista que o provedor opera com uma infraestrutura de comunicação baseada em uma rede de transporte própria de alta velocidade, conforme descrito anteriormente.

A título de ilustração, vejamos qual a infraestrutura da *Huawei Cloud* disponível na China:

País	Localidade	Regiões	Zonas de disponibilidade
China	Karamay	01	01
China	Ulanqab	01	03
China	Pequim	02	07
China	Xangai	01	04
China	Hong Kong	01	04
China	Gulyang	01	03
China	Shenzhen	01	01
China	Guangzhou	01	06
China	Suzhou	01	04

Analisando a tabela, podemos observar que são dez regiões distribuídas em nove localidades distintas do país. No caso de Pequim, existem duas regiões, que normalmente são nomeadas mantendo a identificação da localidade, ou seja, "Pequim", acrescidas da área geográfica, dentro da cidade ou estado em que as infraestruturas se concentram, como Norte, Sul, Leste e Oeste, por exemplo. Por se tratar de uma empresa sediada na China, naturalmente, essa é a localidade em que o provedor possui o maior número de clientes, o que resulta invariavelmente em uma necessidade de maior concentração de recursos frente a outras localidades do globo.

Já na América Latina, a situação é bem diferente, conforme pode ser visto na tabela a seguir:

País	Localidade	Regiões	Zonas de disponibilidade
Argentina	Buenos Aires	01	01
Brasil	São Paulo	01	03
Chile	Santiago	01	02
Peru	Lima	01	02

Cada região possui uma ou mais "zonas de disponibilidade", que se caracterizam, na prática, como agrupamentos de *datacenters* geograficamente localizados dentro da região a que se referem.

As zonas de disponibilidade possuem fornecimento redundante de energia elétrica, sistemas de refrigeração, segurança física perimetral e *links* de conectividade com baixa latência, com vistas à alta disponibilidade dos recursos utilizados. O eventual aumento do número de zonas de disponibilidade em uma região naturalmente está atrelado à demanda de clientes na localidade.

Quanto mais próximo o cliente está da infraestrutura, menor a latência da comunicação e melhor a experiência do usuário. No entanto, definir a região em que se deseja provisionar a infraestrutura de um projeto nem sempre é uma tarefa tão óbvia.

Dentre os fatores que podem impactar o processo de escolha de uma região, destacam-se:

- **Conformidades:** selecionando um entre vários exemplos possíveis, podemos citar que instituições federais brasileiras, por força de lei, só estão autorizadas a utilizar provedores de nuvem que possuam infraestrutura física e representação oficial no país. Tal medida garante que a empresa esteja operando em território nacional e, portanto, seguindo a legislação brasileira, preservando a soberania de dados e a segurança nacional;

- **Custos:** como é do conhecimento de todos, os custos de produtos e serviços, ainda que idênticos, apresentam variações de acordo com a carga tributária de uma cidade, estado ou país. Provisionar infraestrutura na região *Latin America-Brasil-São Paulo* pode ser muito mais oneroso economicamente do que realizar a mesma operação na região *Latin America-Chile-Santiago*, por exemplo;
- **Disponibilidade de serviços:** nem todos os serviços oferecidos pelo provedor de nuvem estão disponíveis globalmente em seu lançamento. Demandas localizadas tendem a estimular o surgimento de novos serviços que, em um primeiro momento, fazem sentido somente em determinadas regiões. A depender do propósito deste, a oferta pode ser estendida para outras, ou todas as regiões, ou ainda permanecer restrita somente às localidades originalmente atendidas.

2.3 Considerações importantes

A cada dia, surgem novas tendências que podem gerar mudanças de impacto significativo, com resultados na agilidade na implantação e manutenção de infraestrutura, na eficiência de integração entre diferentes tecnologias, na otimização de processos cotidianos, dentre outros benefícios que, ao mesmo tempo, oferecem inovações tecnológicas relevantes, sem perder de vista a viabilidade econômica.

Profissionais envolvidos com tecnologias da informação e comunicação precisam tomar decisões estratégicas com relativa frequência, em diferentes segmentos. Para que as escolhas em questão possam ser, em sua grande maioria, assertivas, é uma boa prática incorporar o perfil de estudante incansável, que persegue

cotidianamente o entendimento do panorama geral da área em que escolheu atuar e, inevitavelmente, se especializar.

Nesse sentido, o estudo e a experimentação de tecnologias e soluções de nuvem são fundamentais não somente para dominar os conceitos técnicos da área de atuação específica, mas, em especial, possibilitam ao usuário conhecer alternativas estratégicas que tornam um projeto economicamente viável, de acordo com as condições financeiras e de recursos humanos do cliente.

Questionário online

Figura 2.4: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 2.5: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo

Parte 2: Primeiros passos

Capítulo 3
Precificação

Após entender o conceito e as principais características da tecnologia de computação em nuvem, este é certamente um capítulo que merece toda a sua atenção.

Um(a) profissional atuante na área de computação em nuvem, antes de qualquer coisa, precisa ter em mente que, assim como se deve utilizar tecnologias modernas que promovam agilidade, eficiência e estabilidade no processo de implantação de infraestrutura, é igualmente importante ter a responsabilidade constante de perseguir e dominar as principais estratégias de aquisição de ativos nesse tipo de ambiente.

Esse papel é mandatório para que, em tempos de aumento de casos de repatriação de dados, seja possível defender, com propriedade e responsabilidade, a manutenção da utilização consciente da tecnologia, com vistas à excelência operacional, segurança e otimização de custos. Esse olhar é fundamental tanto para viabilizar novos projetos como para manter os que já estão em funcionamento competitivos e com diferenciais tecnológicos relevantes.

3.1 Estimando o custo de serviços de nuvem

Todos os provedores de nuvem pública disponibilizam aos clientes ferramentas de estimativa de custos, normalmente denominadas "*Price Calculators*". A familiaridade e a efetiva utilização desse tipo de recurso são essenciais, uma vez que é possível não só prever os

custos dos serviços necessários para a implantação de um eventual projeto futuro, como lançar mão de diversas estratégias com vistas à otimização de custos que podem fazer toda a diferença no processo de tomada de decisão.

Atualmente, devemos levar com consideração a estratégia de utilização de ambientes baseados em *multicloud*. A motivação para utilizar mais de um provedor de nuvem pública para provimento de recursos de infraestrutura e serviços, em geral, tem natureza econômica. No entanto, podem ser considerados outros fatores, como alta disponibilidade, balanceamento de carga, recuperação de desastres, redundância, necessidade de utilização de serviços específicos, presença em determinadas regiões do globo, ou ainda, o atendimento a padrões de conformidade ofertados por um número reduzido de provedores, por exemplo.

Nesse sentido, a comparação de custos entre serviços equivalentes ofertados por diferentes fornecedores é crucial, pois disponibiliza diferentes opções de escolha e, em especial, proporciona ao usuário o conhecimento dos pontos fortes e fracos de cada provedor. A seguir, são apresentadas as ferramentas de precificação disponibilizadas pelos principais provedores de nuvem da atualidade:

Provedor	*Website*
Alibaba Cloud	https://www.alibabacloud.com/en/pricing-calculator?/
Amazon Web Services	https://calculator.aws/
Google Cloud Platform	https://cloud.google.com/products/calculator/
Huawei Cloud	https://www.huaweicloud.com/intl/pt-br/pricing/calculator.html#/
IBM Cloud	https://www.ibm.com/cloud/cloud-calculator
Microsoft Azure	https://azure.microsoft.com/en-gb/pricing/calculator/
Oracle Cloud Infrastructure	https://www.oracle.com/br/cloud/costestimator.htn

3.2 Tarifação de serviços básicos

Para um melhor entendimento, vamos conhecer as principais características e funcionalidades da *Price Calculator* da *Huawei Cloud*. No menu disponível no lado esquerdo da imagem, é possível observar as categorias de serviço disponibilizadas pelo provedor, com as respectivas opções disponíveis em cada uma delas. A imagem a seguir exibe a interface de utilização da ferramenta em questão, bem como os parâmetros de configuração do serviço de computação *Elastic Cloud Server* ou, simplesmente, ECS.

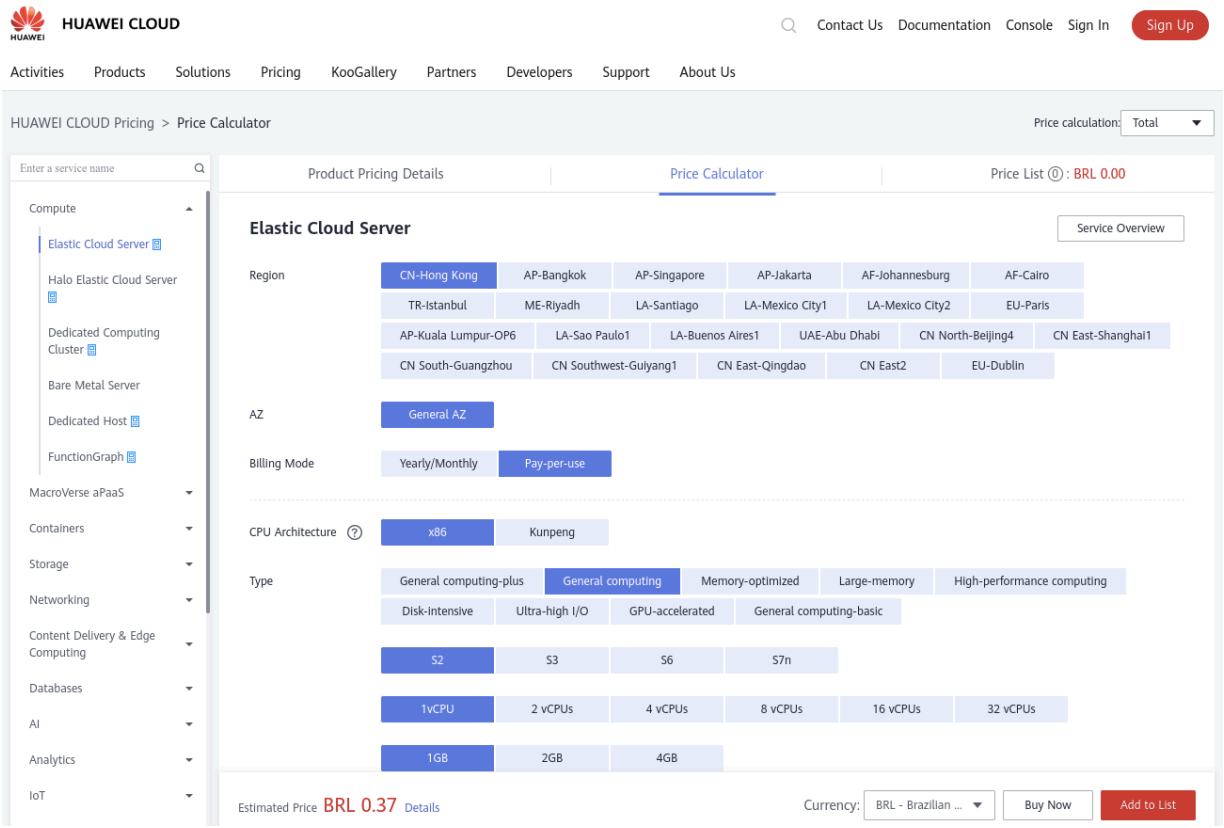


Figura 3.1: Price Calculator da Huawei Cloud

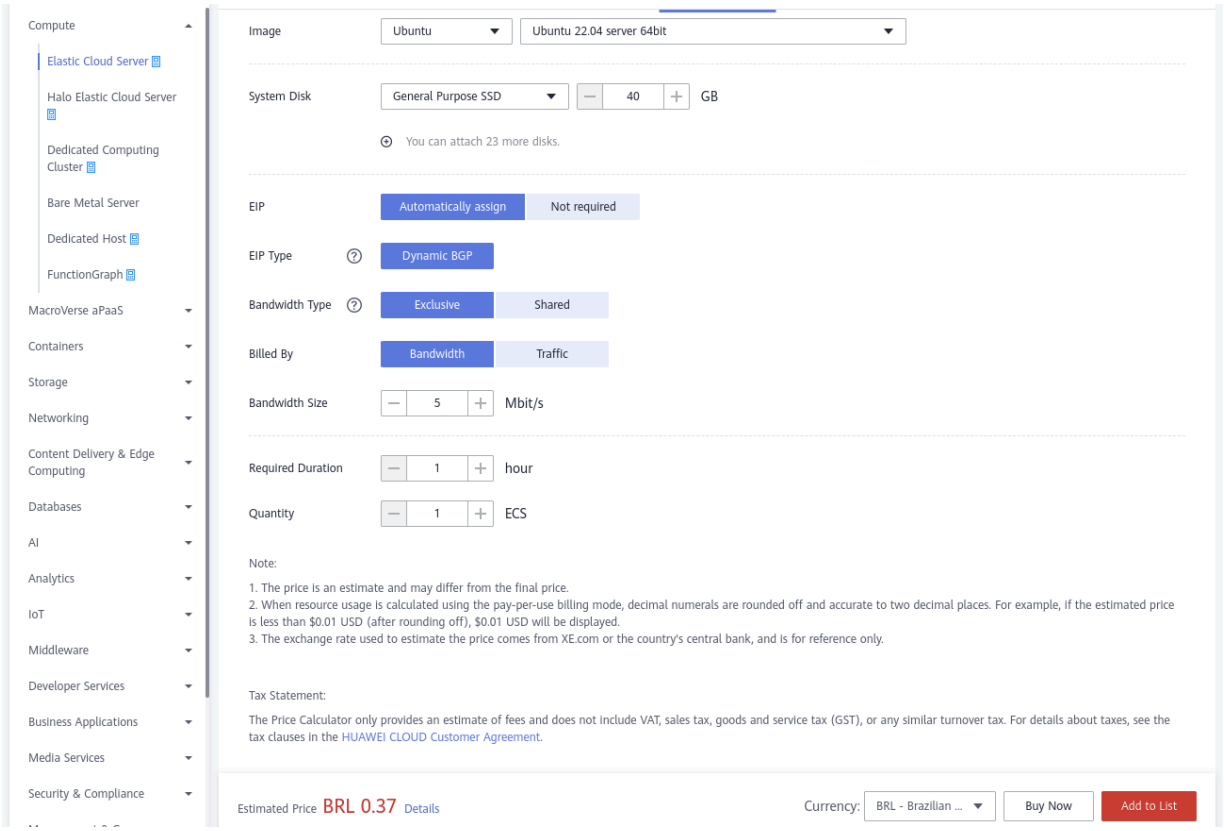


Figura 3.2: Price Calculator da Huawei Cloud

A ferramenta permite que o usuário selecione os parâmetros de configuração e opções de provisionamento que, hipoteticamente, desejaria utilizar inicialmente em seu projeto, a fim de estimar os custos correspondentes. No rodapé da página, à direita, é possível determinar qual a moeda utilizada no processo de precificação. Ainda no rodapé, agora do lado esquerdo e em destaque na cor vermelha, é exibido o custo parcial do serviço atualmente selecionado, informação que é atualizada de acordo com o preenchimento dos campos disponíveis, bem como a modificação dos parâmetros pré-selecionados inicialmente na interface.

Para ilustrar algumas práticas importantes e necessárias no processo de utilização e aprendizagem da ferramenta, vamos utilizar como base de referência uma instância de virtualização com configurações modestas, equivalentes a uma máquina de entrada utilizada somente para prova de conceito, conforme descrito a seguir:

Parâmetro	Valor
Região	China, Hong Kong
Zona de disponibilidade	Selecionada de forma aleatória
Modo de pagamento	Conforme o uso
Arquitetura de CPU	x86
Tipo	Computação de propósito geral
vCPU	01
Memória RAM	1 GB
Instância	s2.small.1
Imagem	Ubuntu Server
Dispositivo de Armazenamento	SSD de propósito geral
Capacidade	40 GB
Elastic IP	Não utilizado
Período	01 hora
Quantidade	01 ECS
Preço	USD 0,02 por hora

A configuração definida inicialmente corresponde ao custo médio estimado de USD 0,02 por hora, na modalidade de pagamento denominada *"pay-per-use"*. Considerando que o dia possui 24 horas, o custo diário da instância corresponde a USD 0,41. Em um mês completo, composto por 31 dias ou 744 horas, o custo mensal estimado do serviço é de USD 12,56. Em um ano, com 12 meses ou 8.928 horas, seriam gastos aproximadamente USD 150,70 com a instância de virtualização em questão.

O tempo de utilização do recurso considerado no processo de estimativa de precificação é considerado em horas e pode ser modificado diretamente no campo *"Required Duration"*, localizado como penúltimo item no final da página. É importante lembrar que os valores apresentados foram todos obtidos a partir dos cálculos realizados pela ferramenta. Nesse sentido, certamente ocorrem processos de arredondamento e uso de casas decimais adicionais, omitidas da interface para efeito de formatação, entre outros fatores que eventualmente vão produzir valores ligeiramente diferentes daqueles que seriam obtidos por meio de uma calculadora convencional. Os valores estimados na primeira simulação são apresentados na tabela a seguir.

Período	Quantidade de horas	Valor em USD
Hora	01	0,02
Dia	24	0,41
Mês	744	12,56
Ano	8928	150,70

A contratação na modalidade *"pay-per-use"* utiliza valores denominados "de tabela cheia" e é indicada para clientes com cargas de trabalho sujeitas a picos de utilização imprevisíveis em um primeiro momento, ou ainda para aqueles que não desejam assumir compromissos de contratação de médio ou longo prazo firmados via contratos. Clientes que queiram economizar dinheiro podem optar pelas opções de pagamento mensal ou anual que, a depender do serviço, podem disponibilizar descontos interessantes. No caso da assinatura anual, a economia proporcionada no exemplo de referência foi de USD 25,40, um desconto de aproximadamente 20%.

Período em meses	Valor em USD
01	12,70
12	127,00

Importante destacar que as despesas dos chamados "serviços básicos" são complementares. Para a composição de valores do exemplo supracitado, são considerados os custos do serviço de computação (USD 77,00), do serviço de armazenamento de blocos, de acordo com a capacidade escolhida (40 GB = USD 50,00) e, proporcionalmente, do sistema operacional utilizado, que, no exemplo em questão, é nulo por se tratar de um *software* com licença *GPL*. O detalhamento da cobrança também nos permite verificar o valor do desconto aplicado, em função da mudança do *billing mode* para a modalidade *yearly/monthly*.

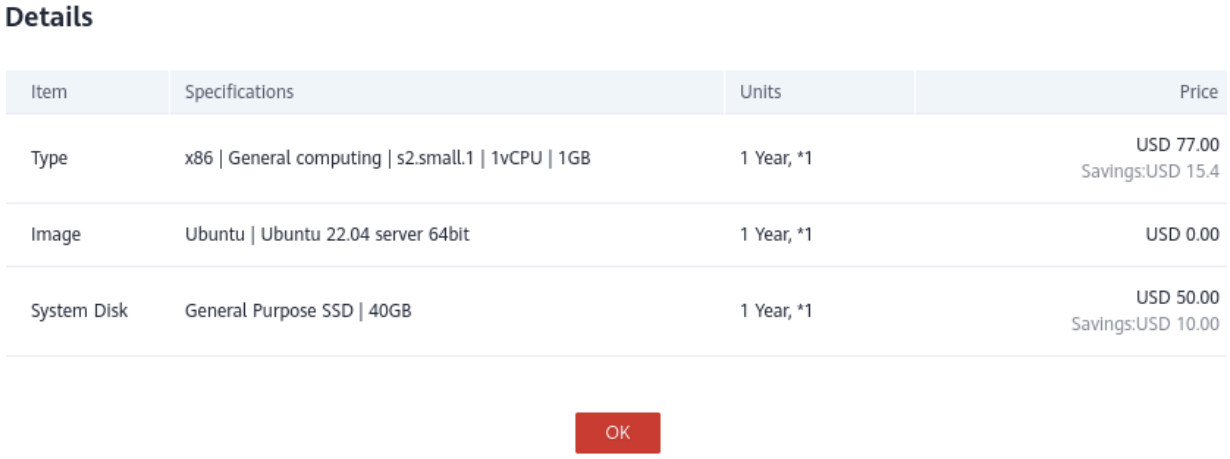


Figura 3.3: Detalhes da cobrança do plano anual com os respectivos descontos em abril de 2025

A utilização de recursos adicionais não obrigatórios, como o provimento de um endereço IP elástico, impacta significativamente nos custos da instância. Retomando a mesma configuração utilizada inicialmente, foi adicionado um *Elastic IP* com largura de banda exclusiva de 5 *Mbit/s*, conforme a seguir:

Parâmetro	Valor
Região	China, Hong Kong
Zona de disponibilidade	Selecionada de forma aleatória
Modo de pagamento	Conforme o uso
Arquitetura de CPU	x86
Tipo	Computação de propósito geral
vCPU	01
Memória RAM	1 GB
Instância	s2.small.1
Imagem	Ubuntu Server
Dispositivo de Armazenamento	SSD de propósito geral
Capacidade	40 GB
Elastic IP	Atribuído automaticamente
Elastic IP - Tipo	BGP dinâmico
Tipo de Largura de Banda	Exclusiva
Tarifação	Tráfego
Largura de Banda	5 Mbit/s
Período	01 hora
Quantidade	01 ECS
Preço	USD 0,07 por hora

Nesse cenário, passaram a vigorar os seguintes valores:

Período	Quantidade de horas	Valor em USD
Hora	01	0,07
Dia	24	1,61
Mês	744	49,76
Ano	8928	597,10

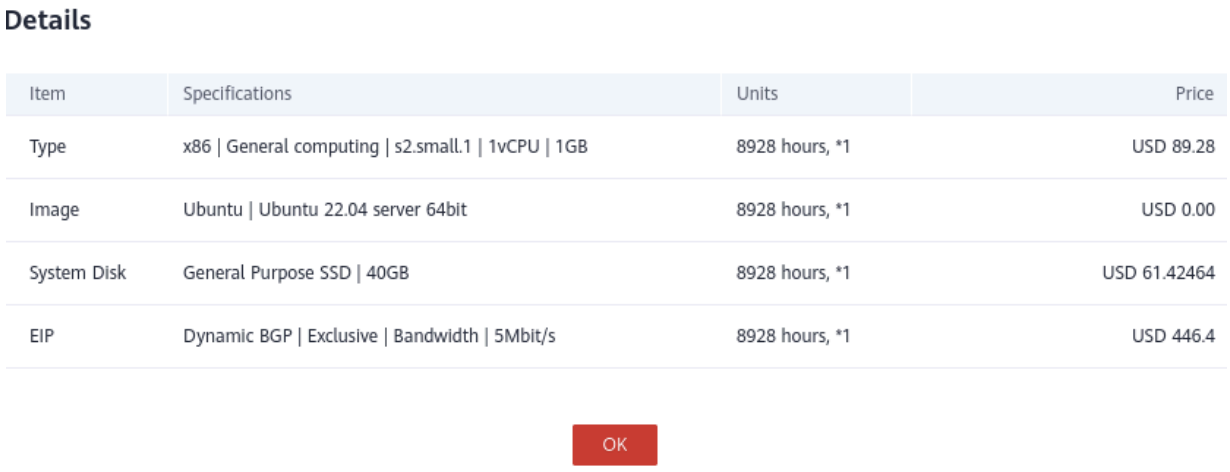


Figura 3.4: Detalhes da cobrança com inclusão de um Elastic IP com largura de banda exclusiva de 5 Mbit/s, na modalidade "pay-per-use"

Conforme dito anteriormente, em alguns casos, os critérios para determinação da região em que a infraestrutura deve ser provisionada não dependem apenas do fator econômico, mas também do atendimento a padrões de conformidade ou normativas que devem ser implementadas por força de lei. Instituições federais brasileiras, por exemplo, estão restritas à utilização de provedores de nuvem que possuam regiões disponibilizadas no país, tendo em vista que eles devem obrigatoriamente seguir as leis do Brasil. Nesse caso específico, caso o provedor não possuísse oferta de serviços localmente, ele não poderia ser utilizado. O impacto financeiro da mudança em questão é significativo. Mantendo-se as mesmas configurações de referência originalmente selecionadas, sem *Elastic*

IP, mas com provisionamento de infraestrutura na região *Latin America, São Paulo 01*, os custos mais que triplicam, passando a representar um custo de USD 0,07 por hora, conforme a seguir:

Período	Quantidade de horas	Valor em USD
Hora	01	0,07
Dia	24	1,73
Mês	744	53,48
Ano	8928	641,74

Tendo em vista que a escolha de provisionamento da infraestrutura em uma região fora do país não é uma opção para o caso apresentado, cabe ao administrador buscar alternativas para tornar o projeto economicamente viável. Na região em questão, existe a possibilidade de adquirir instâncias reservadas, que permitem descontos substanciais mediante assinatura de contratos de 12 ou 36 meses, conforme a seguir.

Período em anos	Valor em USD
01	131,40
03	262,80

Para efeito de comparação, a simulação apresentada a seguir considera as mesmas configurações de referência, com provisionamento de infraestrutura na região *Latin America, Santiago*, com localização geográfica relativamente próxima à região *Latin America, São Paulo 01*, que seria a escolha natural.

Período	Quantidade de horas	Valor em USD
Hora	01	0,02
Dia	24	0,46
Mês	744	14,28
Ano	8928	171,42

Caso a aplicação em questão tenha a latência como um fator sensível, evidentemente a solução mais óbvia seria provisionar a infraestrutura no país em que o público-alvo está localizado. No entanto, na grande maioria dos casos, o fator econômico desempenha papel determinante no processo de tomada de decisão. Nesse sentido, soluções paliativas que atendam à necessidade do negócio e possuam uma relação custo-benefício atrativa tendem a se tornar escolhas mais inteligentes.

Na região supracitada, a possibilidade de adquirir instâncias reservadas segue disponível nos planos de contratação de 12 ou 36 meses, tornando a alternativa economicamente ainda mais interessante. O valor anual praticado no contrato de 12 meses corresponde a um desconto de aproximadamente 55% em relação à opção equivalente na modalidade "*pay-per-use*". Uma oportunidade atrativa é oferecida na aquisição do plano de 36 meses. O valor praticado em contrato é aproximadamente 8% menor quando comparado ao valor anual praticado na modalidade "*pay-per-use*".

Período em anos	Valor em USD
01	78,84
03	157,68

Vale lembrar que, apesar de apresentar valores mais atrativos e muito próximos aos obtidos na região *China – Hong-Kong*, a alternativa apresentada no estudo de caso só será viável caso o

projeto não exija conformidade legal que determine que a infraestrutura esteja localizada no mesmo país da empresa responsável.

Outro requisito obrigatório é a realização prévia de testes de *benchmark* de latência, a fim de assegurar que os requisitos de desempenho do projeto sejam atendidos.

3.3 Considerações importantes

As sugestões de regiões, configurações de equipamentos, modelos de contratação e os respectivos valores utilizados nos exemplos de precificação apresentados neste capítulo devem ser considerados apenas como recursos didáticos necessários para o entendimento dos conceitos abordados.

As possibilidades de simulação disponibilizadas pela ferramenta permitem ao usuário combinar diferentes fatores com vistas ao atendimento de necessidades de projeto e à viabilização de oportunidades economicamente atrativas, que podem fazer toda a diferença no processo de tomada de decisão em um eventual projeto de implementação de infraestruturas em ambiente de nuvem pública.

O portfólio de serviços oferecido pelos provedores de nuvem pública, em geral, é significativamente amplo. Nesse sentido, recomendamos a você que retome a utilização da ferramenta de precificação ao final de cada capítulo concluído.

Essa boa prática possibilita um melhor entendimento da *"Price Calculator¨*, considerando suas particularidades e, em especial, os artifícios que podem ser aplicados em cada uma das categorias de serviço, com vistas à obtenção de tarifas competitivas e diferenciais de performance. Ao mesmo tempo, possibilita a aplicação e a fixação dos conhecimentos específicos adquiridos em cada um dos capítulos.

Questionário online

Figura 3.5: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 3.6: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo

Capítulo 4
Serviços de autenticação, autorização e auditoria

O primeiro passo necessário para a utilização de um provedor de nuvem pública é a criação de uma conta. Todos os recursos que eventualmente venham a ser criados na infraestrutura de nuvem estarão sob o "guarda-chuva" dessa conta que, em geral, é denominada como "conta *root*".

Em um primeiro momento, os parâmetros solicitados para o registro da conta são básicos, restringindo-se ao país de origem, a uma conta de *e-mail* válida, à senha desejada e a um código de confirmação enviado ao *e-mail* informado, utilizado como segundo fator de autenticação no processo de registro. A criação de uma nova conta na *Huawei Cloud* pode ser feita em: https://reg.huaweicloud.com/registerui/intl/register.html?locale=pt-br.

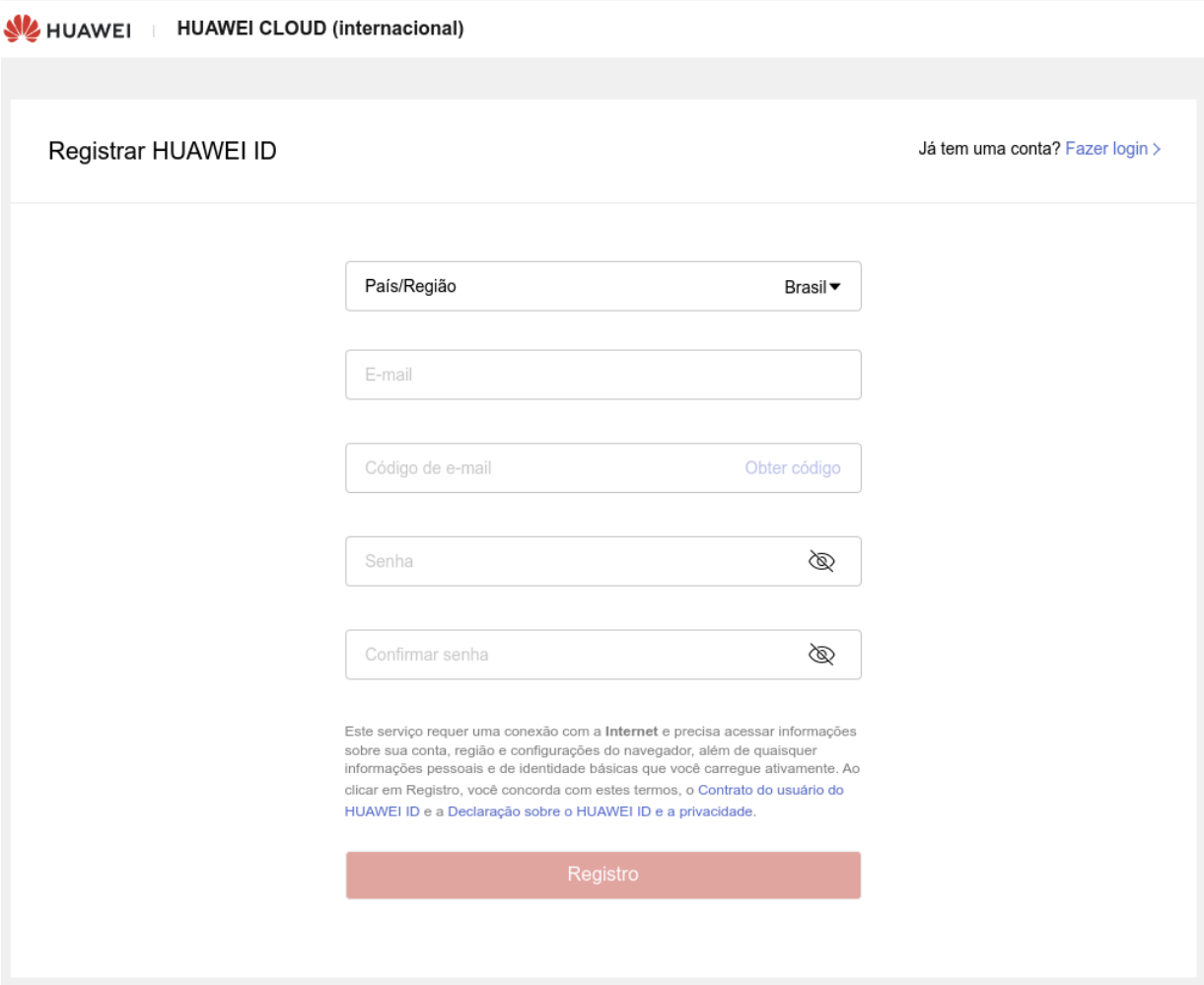


Figura 4.1: Registro de uma nova conta na Huawei Cloud

Finalizado o processo de registro da nova conta, o cliente deverá proceder com o primeiro acesso à plataforma, processo denominado *autenticação*. Nessa etapa, o usuário deve provar que é quem afirma ser. Para tanto, deve disponibilizar informações, credenciais ou atributos por meio do teclado ou da utilização de dispositivos físicos ou biométricos que comprovem sua identidade. No caso de uma conta *root*, as credenciais de acesso são baseadas em *e-mail* e senha, respectivamente.

Figura 4.2: Interface de login da Huawei Cloud

Uma vez concluído com sucesso o processo de autenticação, o usuário passa a ter acesso a todos os recursos e serviços disponibilizados pelo provedor de acordo com os privilégios concedidos ao perfil a que está associado. Essa etapa recebe o nome de *autorização*. Como estamos falando de uma conta *root*, o usuário em questão não possui nenhum tipo de bloqueio ou restrição de uso, o que torna sua utilização cotidiana um risco.

4.1 Interface de gerenciamento do serviço

O console de gerenciamento *web* é o ponto de partida do usuário para a utilização de produtos e serviços disponibilizados pelo provedor de nuvem pública. A interface exibida a seguir fica disponível ao usuário logo após o processo de *login*.

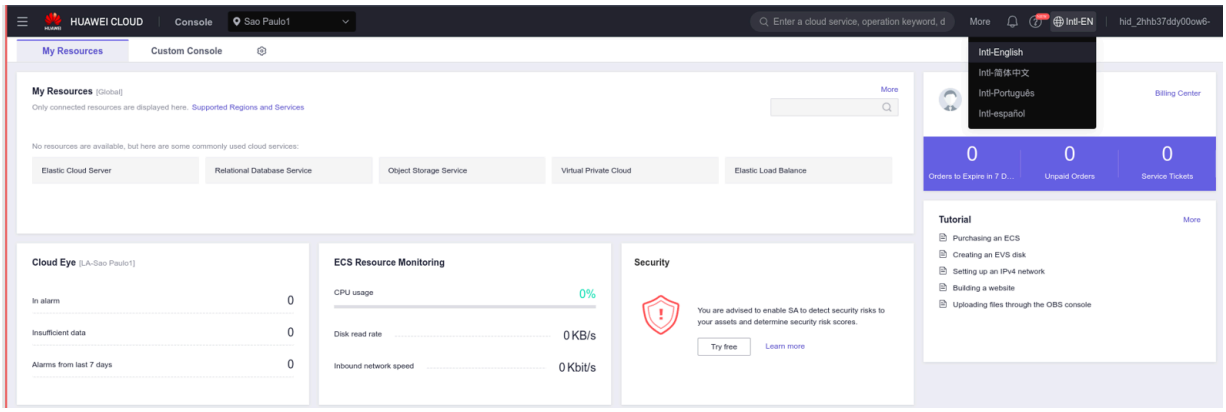


Figura 4.3: Console de gerenciamento web da Huawei Cloud

Na imagem, é possível observar que o console de gerenciamento *web* é um *dashboard* que concentra as informações mais relevantes relacionadas aos recursos atrelados à conta em questão. Como se trata de uma conta recém-criada, não existem, obviamente, informações a serem listadas nos painéis de recursos (*My Resources*), de monitoramento de alarmes (*Cloud Eye*) ou de monitoramento de recursos de virtualização (*ECS Resource Monitoring*).

Dentre outras informações relevantes, podemos observar a região utilizada pelo usuário no momento — neste caso, "São Paulo 1" — bem como os idiomas disponíveis para internacionalização da interface com o usuário. O *layout* do painel de controle exibido inicialmente é comum a todo novo usuário. No entanto, a disposição e as informações contidas em cada painel, assim como a eventual necessidade de exibir diferentes métricas — as quais, a princípio, não existem —, variam de acordo com a demanda de cada usuário. Desse modo, a interface padrão do console de gerenciamento *web*

pode ser customizada de acordo com as necessidades do cliente por meio da opção "*Custom Console*".

As boas práticas de segurança recomendam que políticas de acesso sejam definidas com base no princípio do privilégio mínimo. Em geral, contas estão associadas a identidades que representam usuários ou aplicativos. Dessa forma, é essencial que a entidade autenticada tenha acesso apenas aos recursos e funcionalidades absolutamente necessários para que consiga desempenhar o seu papel no sistema.

Para que isso seja possível, recomenda-se que, imediatamente após o primeiro acesso à conta associada ao usuário *root*, seja prontamente criado um usuário com perfil administrativo para que ele passe a lidar com as tarefas cotidianas associadas a esse tipo de demanda. A conta *root*, embora funcional, deve ser reservada para utilizações pontuais relacionadas a mudanças de informações de cobrança, planos de suporte e outras atribuições específicas que variam de provedor para provedor, mas que ocorrem com pouca frequência. Essa prática é a mesma recomendada em sistemas operacionais convencionais: usuários não privilegiados são utilizados para as tarefas rotineiras e o usuário *root* realiza tarefas administrativas pontuais somente quando necessário.

Uma vez autenticado em sua conta *root*, o cliente tem acesso a todo o portfólio de produtos e serviços disponibilizados pelo provedor de nuvem via console de gerenciamento *web*. Para que o usuário com perfil administrativo possa ser criado, faremos uso do serviço denominado "*Identity and Access Management*", ou simplesmente *IAM*. Para tanto, basta digitar na barra de buscas o nome do serviço de nuvem desejado conforme a seguir.

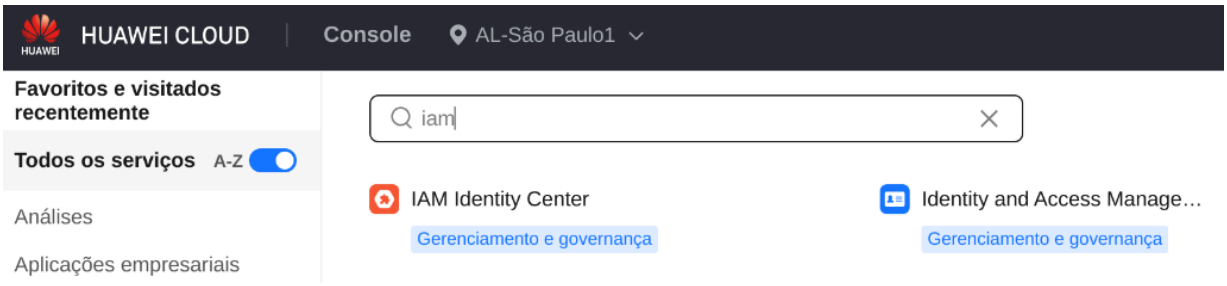


Figura 4.4: Pesquisando serviços de nuvem disponibilizados pelo provedor

Por se tratar de um serviço básico, necessário para utilização de todos os demais, os principais provedores de nuvem do mercado oferecem serviços de autenticação e autorização com a mesma finalidade e com nomenclatura semelhante ou idêntica.

Provedor	***Website***
Alibaba Cloud	https://www.alibabacloud.com/en/product/identity-as-a-service-idaas?/
Amazon Web Services	https://aws.amazon.com/pt/iam/
Google Cloud Platform	https://cloud.google.com/security/products/iam?hl=pt-br
Huawei Cloud	https://www.huaweicloud.com/intl/pt-br/product/iam.html
IBM Cloud	https://www.ibm.com/br-pt/products/verify-saas/cloud
Microsoft Azure	https://azure.microsoft.com/pt-br/products/category/identity
Oracle Cloud Infrastructure	https://www.oracle.com/br/security/cloud-security/identity-cloud/

4.2 Gerenciamento de usuários

Ao ingressar no serviço *IAM*, a pessoa gestora da conta tem a possibilidade de verificar quais são os usuários existentes para utilização. Como se trata de uma conta nova, somente o usuário *root* está disponível. Ao contrário do que se poderia imaginar, o *username* associado ao usuário em questão não é *root*, mas sim uma identidade gerada randomicamente no processo de configuração do perfil. Esse mesmo identificador também é utilizado para identificar o nome da conta.

Grande parte dos provedores permite substituir a identificação do usuário root — e, consequentemente, da conta — por um *alias* mais amigável. Essa é uma operação recomendada, pois, como veremos mais adiante, novos usuários adicionados à conta utilizarão um terceiro parâmetro, baseado na identificação da conta, para realizar o processo de *login* no console de gerenciamento *web*. Usar um identificador complexo que não faça sentido para o usuário pode provocar erros desnecessários no processo de autenticação.

Para alterar o *Login ID* atrelado ao usuário, basta posicionar a seta do cursor do dispositivo apontador em cima do nome inicialmente atribuído, informação localizada no canto superior direito da tela. Em seguida, clique na opção "*Basic Information*". Na interface exibida a seguir, ao lado do parâmetro "*HUAWEI ID Information*", clique em "*Manage*".

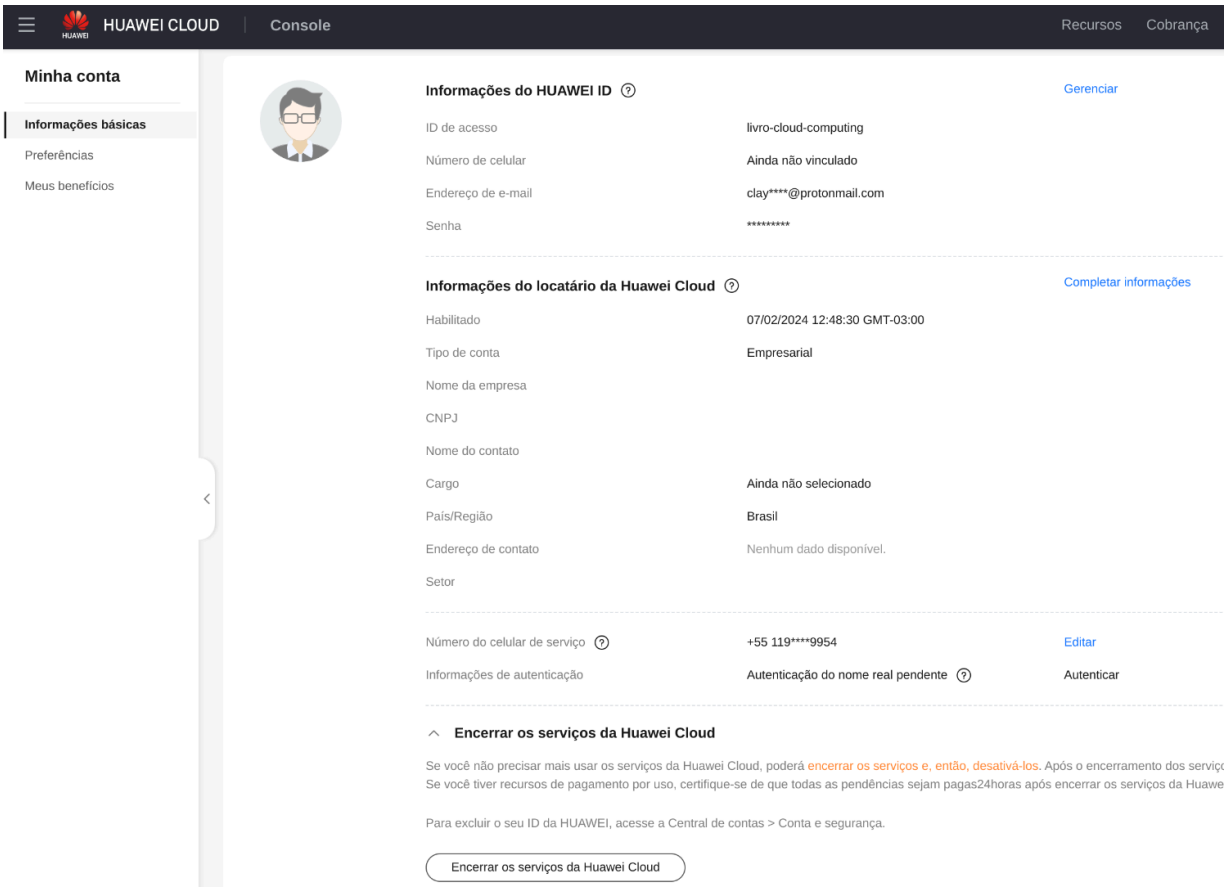


Figura 4.5: Alteração do Login ID da conta - Passo 01/02

Posteriormente, será exibida uma nova interface; basta clicar no botão "*CHANGE*" para que um novo *Login ID* seja definido. Uma vez modificada essa informação, não será mais possível alterá-la. No exemplo em tela, vamos alterar o nome da conta para "livro-cloud-computing". Recomenda-se que, imediatamente após efetivar a mudança, seja realizado um novo *login* na conta *root* para validar as informações modificadas.

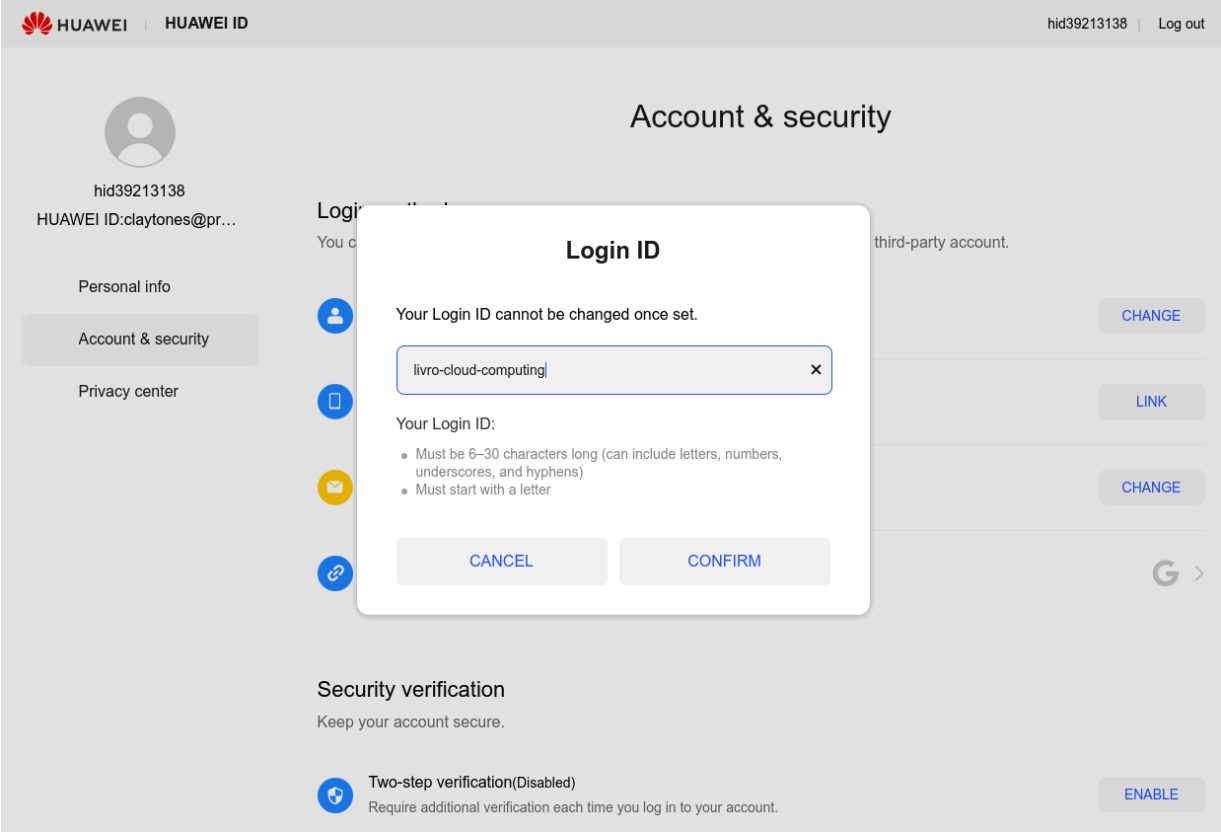


Figura 4.6: Alteração do Login ID da conta - Passo 02/02

Validado o processo de autenticação, podemos retornar ao gerenciamento de usuários dentro do *IAM* e verificar a mudança do *username* associado à conta *root* no quadro de usuários disponíveis, bem como a identificação de nomenclatura exibida no canto superior direito da tela.

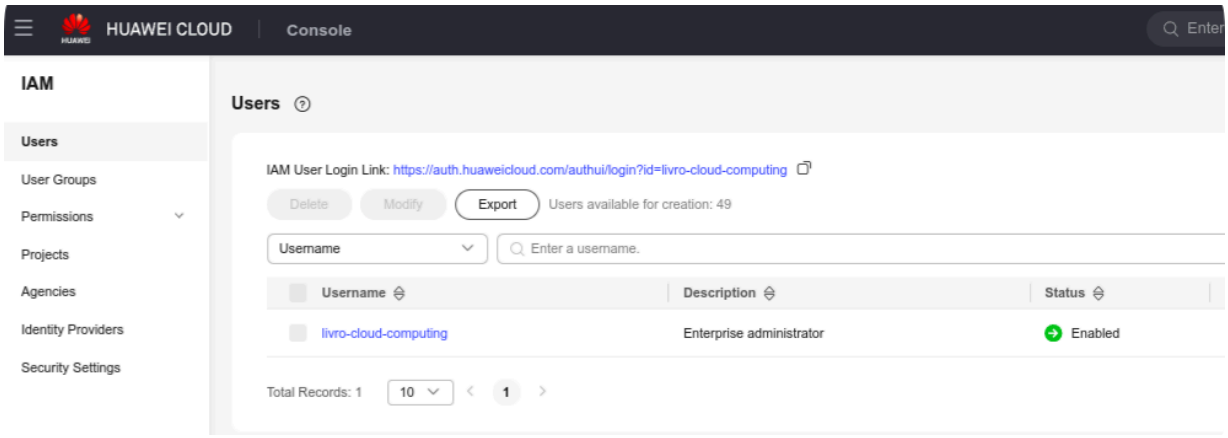


Figura 4.7: Visualização dos usuários disponíveis para utilização

Para criarmos um novo usuário denominado "administrador", clique no botão "*Create User*" localizado no topo superior direito da tela, logo abaixo do nome que identifica a conta. Ao fazê-lo, inicia-se o processo de criação de usuário, composto por três etapas:

- A primeira delas consiste no fornecimento de informações típicas relacionadas ao usuário. Em seguida, define-se o perfil de acesso. O acesso programático deve ser utilizado por aplicativos ou *scripts* que vão consumir os recursos provisionados em nuvem via *APIs*, *CLI* ou *SDKs*; já o acesso via console, como o próprio nome sugere, é voltado aos usuários que vão gerenciar e monitorar recursos via console de gerenciamento *web*. Caso faça sentido, é possível ainda ativar os dois perfis de acesso para um mesmo usuário. Usuários com perfil programático requerem uma *access key*, ao passo que usuários com acesso via console de gerenciamento precisam de uma senha, que pode ser definida de forma customizada ou gerada automaticamente já na interface de criação do usuário ou, ainda, definida pelo usuário via *link* enviado para o e-mail indicado em *user details*. Por fim, recomenda-se ativar um segundo fator de autenticação — via *SMS*, *e-mail* ou *software token* — para adicionar uma camada extra de proteção ao processo de autenticação do usuário em questão.

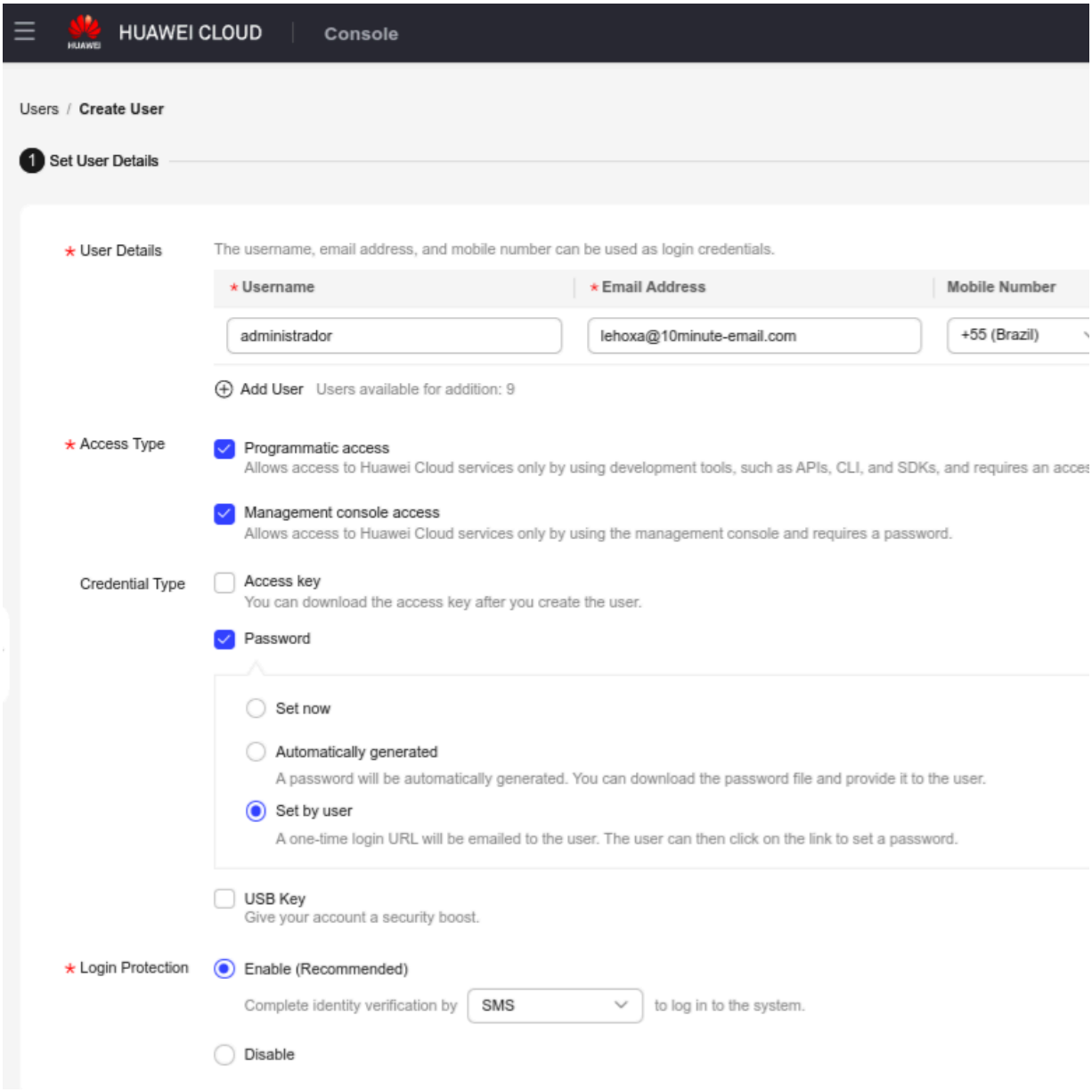


Figura 4.8: Processo de criação de um novo usuário - Etapa 01/03

- Na segunda etapa, o usuário a ser criado pode ser adicionado a um grupo preexistente, ou a um novo grupo, que pode ser criado clicando no *link "create new groups"* disponibilizado acima da caixa "*User Group Name/Description*". Apesar de se tratar de um passo opcional, adicionar um novo usuário a um grupo que contenha outros com o mesmo perfil de uso é altamente recomendado, pois tarefas de operação e

manutenção cotidianas envolvendo a gestão de grupos tornam-se muito mais produtivas e menos propensas a erros, frente à gestão de usuários realizada de forma individual.

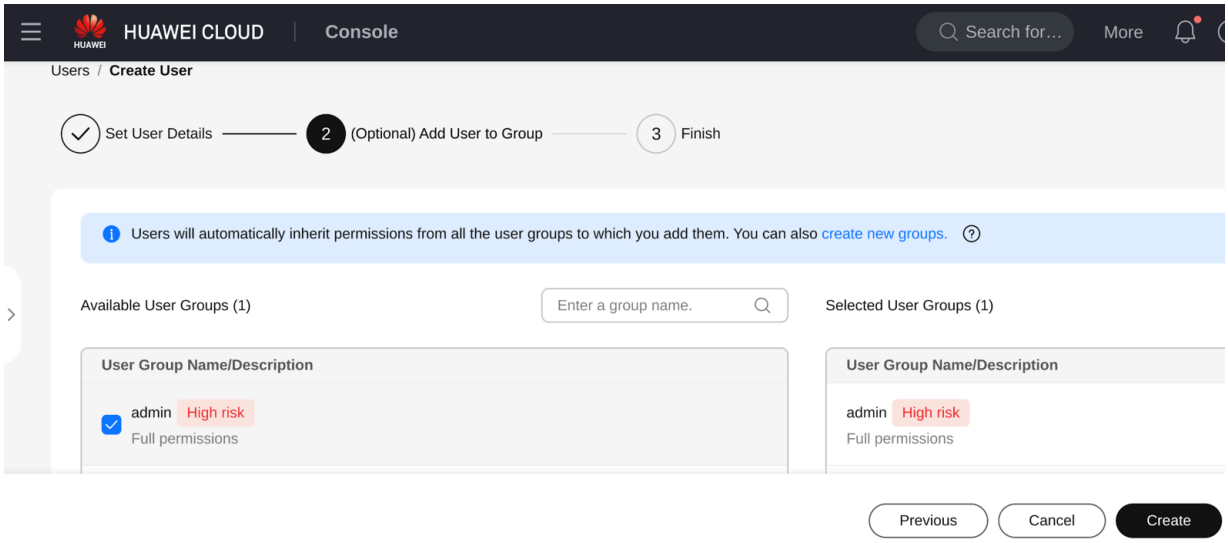


Figura 4.9: Processo de criação de um novo usuário - Etapa 02/03

- Por fim, são exibidas as informações referentes ao usuário recém-criado, na terceira e última etapa referente ao processo.

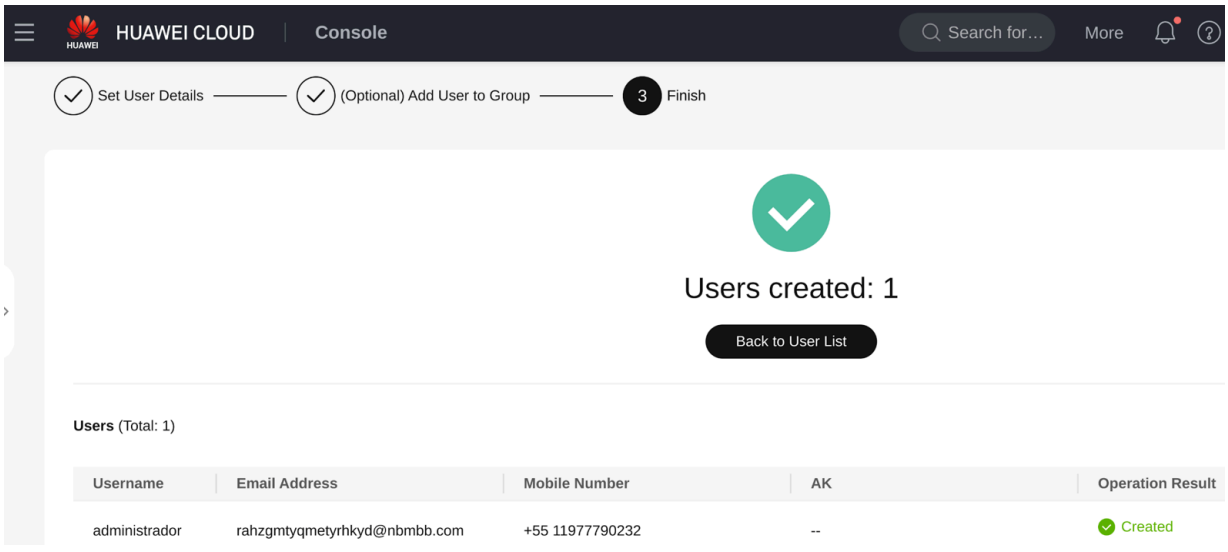


Figura 4.10: Processo de criação de um novo usuário - Etapa 03/03

Uma vez concluído o processo, é necessário realizar o *logout* da conta atual e fazer *login* utilizando as novas credenciais de acesso referentes ao usuário criado. Para tanto, devemos checar a caixa de entrada da conta do *e-mail* associada ao usuário e clicar no *link* disponibilizado. Vale lembrar que essa opção não é a padrão, mas tende a ser uma das mais usuais em ambientes de produção.

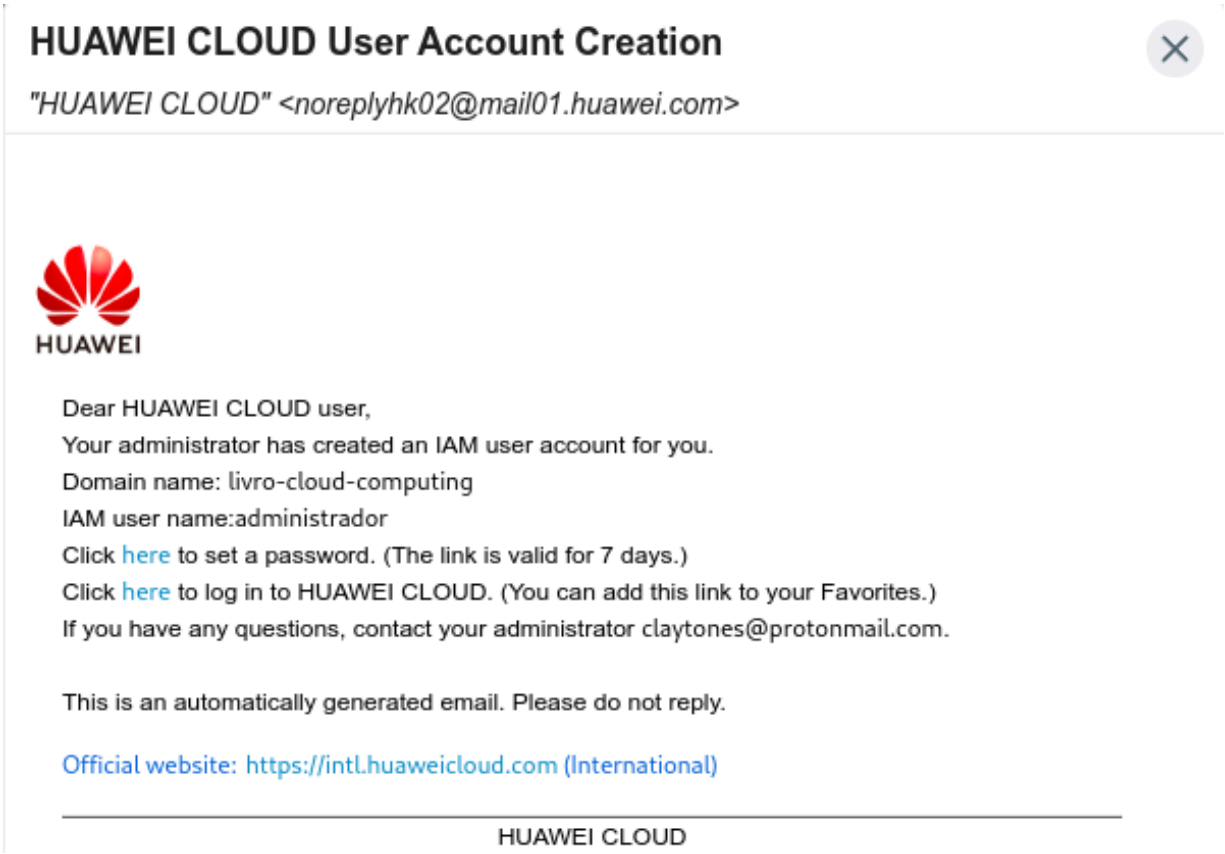


Figura 4.11: E-mail com procedimento para criação de senha de acesso enviado para o novo usuário

Após clicar no endereço informado por *e-mail*, o usuário será direcionado a uma página onde deverá informar a senha desejada, de acordo com a política de senhas vigente do provedor de nuvem. Definida a senha, cabe ao usuário confirmá-la no campo seguinte. Uma vez validadas como equivalentes, o processo será concluído.

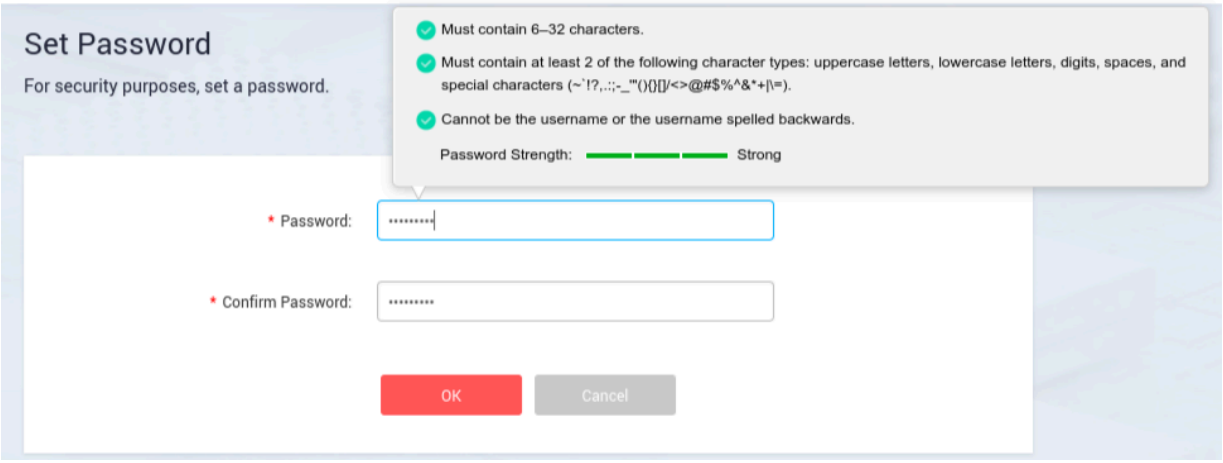


Figura 4.12: Definição da senha do usuário administrador

Enfim, o usuário poderá ser utilizado. Para isso, é necessário realizar um novo *login* no portal de autenticação do provedor de nuvem. A partir de agora, os acessos baseados em *e-mail* e senha, típicos de um acesso referente à conta *root*, não serão mais utilizados. Em seu lugar, vamos selecionar o método de autenticação "*IAM User Login*", localizado abaixo do botão de "*Login*". Nessa modalidade de acesso, além dos campos de usuário e senha, solicita-se a identificação da conta à qual o usuário está atrelado, daí a importância da utilização de um nome significativo, como discutido anteriormente.

A imagem a seguir ilustra o preenchimento dos campos utilizando a conta "livro-cloud-computing" e o usuário "administrador", conforme os parâmetros informados nas etapas anteriores.

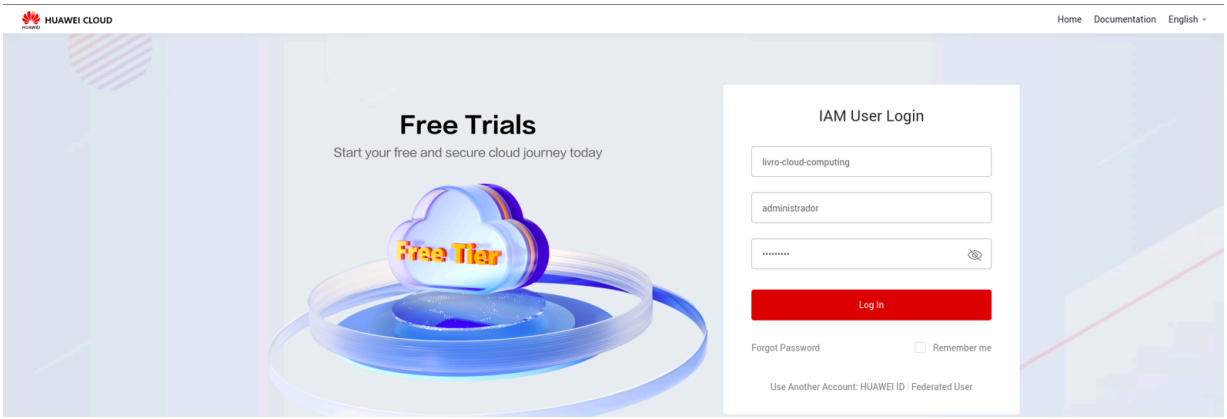


Figura 4.13: Processo de Login utilizando usuário IAM

Uma vez validados os dados no processo de autenticação, o usuário administrador é autorizado a utilizar os recursos pertinentes à sua conta, conforme a interface exibida abaixo:

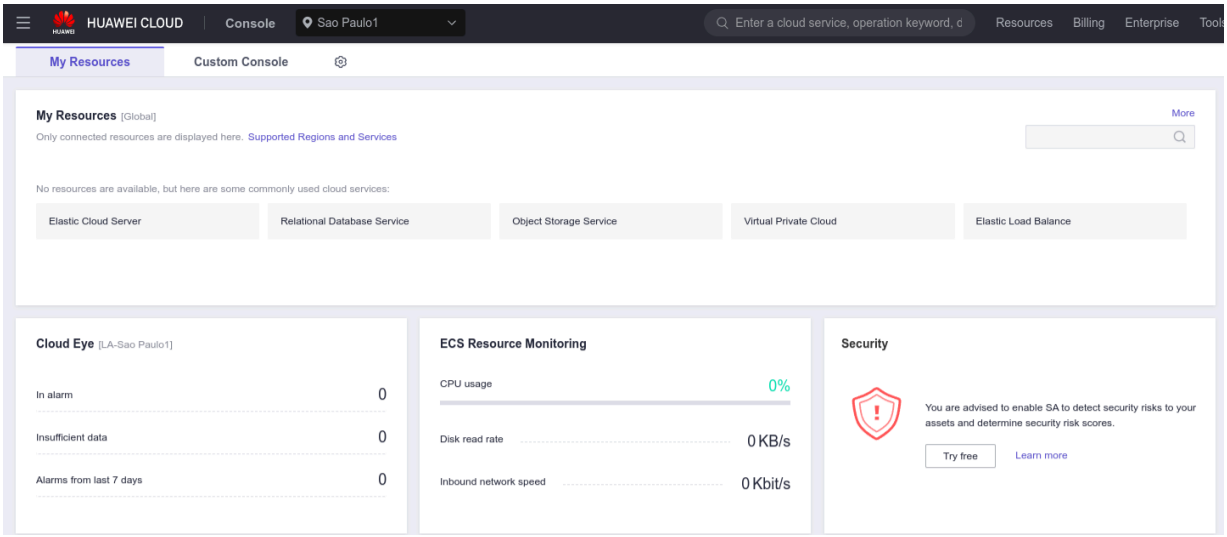


Figura 4.14: Painel de controle do administrador

Ao retornarmos ao menu principal do serviço *IAM*, é possível visualizar os dados referentes aos usuários válidos devidamente atualizados na interface, que agora passa a contar com duas entradas distintas.

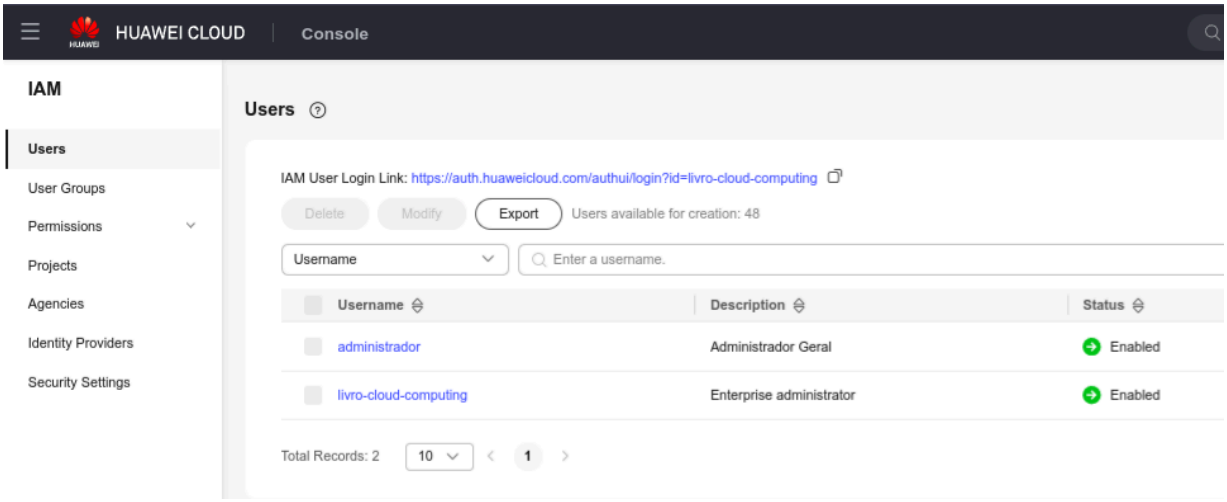


Figura 4.15: Visualização dos usuários disponíveis - Atualização

4.3 Gerenciamento de grupos e definição de políticas de uso

Os exemplos apresentados na seção anterior devem ser considerados como pré-requisitos fortemente recomendados a qualquer usuário responsável por uma conta em um provedor de nuvem pública. É importante ter em mente que o usuário "administrador" pertence ao grupo "admin", o qual, por sua vez, é portador de "*Full permissions*", ou seja, possui alto grau de privilégios administrativos. Portanto, parece óbvio que a estratégia utilizada anteriormente deve ser repetida inúmeras vezes, à medida que o administrador delegar responsabilidades a usuários específicos, sempre obedecendo o princípio do privilégio mínimo.

Para ilustrar essa linha de raciocínio, criamos um usuário chamado "clayton", utilizando os mesmos passos descritos na seção anterior; no entanto, sem vínculo a nenhum grupo. O usuário em questão foi criado, a princípio, com a descrição "Time de Operações", pois estamos, hipoteticamente, falando de um profissional com qualificações na área de redes e infraestrutura.

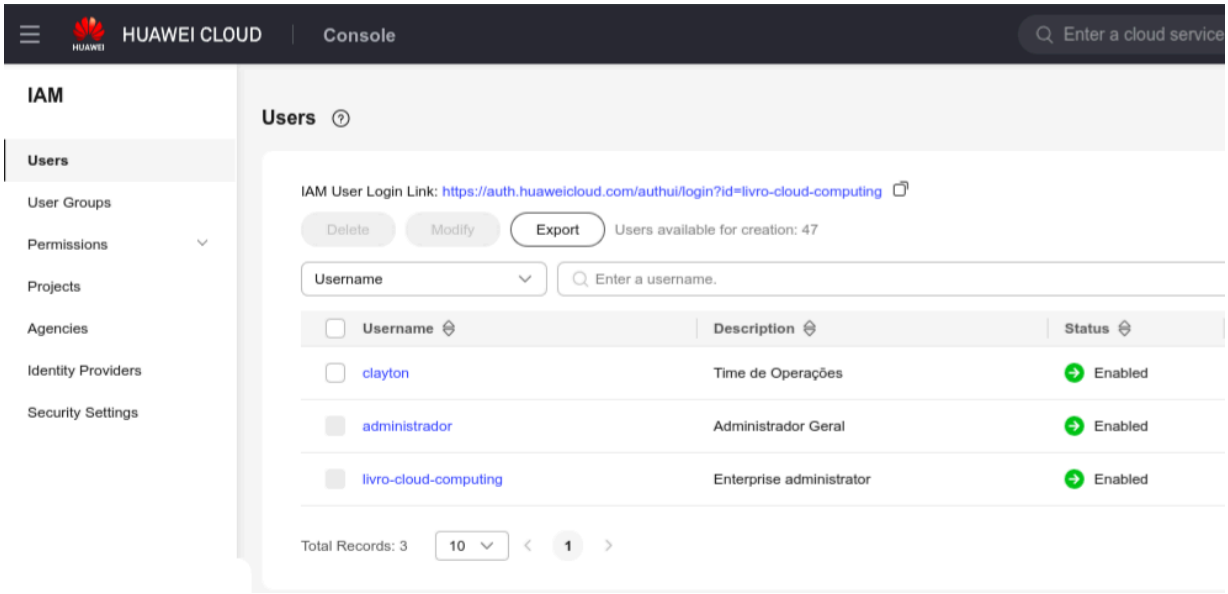


Figura 4.16: Novo usuário "clayton" adicionado

Assim como o usuário recém-criado, certamente teremos mais profissionais que vão compor o time de operações. Dessa forma, é uma boa prática criarmos um grupo que agrupará todos os novos usuários que, eventualmente, tenham esse tipo de competência profissional. O nome escolhido para o grupo é "infraestrutura".

Para criarmos um novo grupo no serviço *IAM* da *Huawei Cloud*, basta selecionar, no menu à esquerda, a opção "*User Groups*" e clicar no botão "*Create User Group*", localizado no canto superior direito da tela, logo abaixo da identificação da conta e do usuário logados no console *web*. A ferramenta vai solicitar um nome para o grupo a ser criado e uma descrição que, apesar de não ser obrigatória, é fortemente recomendada. Uma vez preenchidas as informações solicitadas, basta concluir a operação clicando no botão "OK".

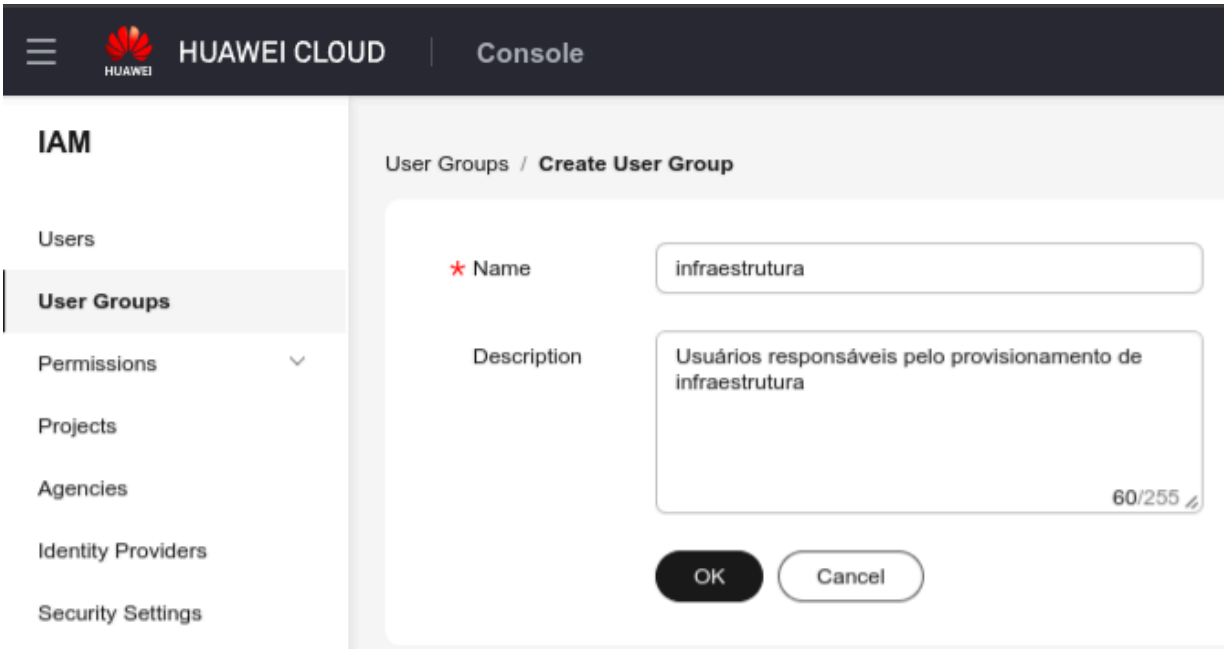


Figura 4.17: Processo de criação de um novo grupo de usuários

Uma vez criado o novo grupo de usuários, vamos agora vincular usuários preexistentes a ele. Para realizar essa operação, vamos clicar na opção "*Manage User*", localizada à direita do grupo desejado, na opção "*User Groups*" do *IAM*.

Figura 4.18: Processo de gerência de grupos de usuários

A interface exibida permite que o operador selecione os usuários desejados para que estes sejam adicionados ao grupo em questão. No nosso caso, selecionaremos o usuário "clayton", criado previamente. Em seguida, clicaremos no botão "OK".

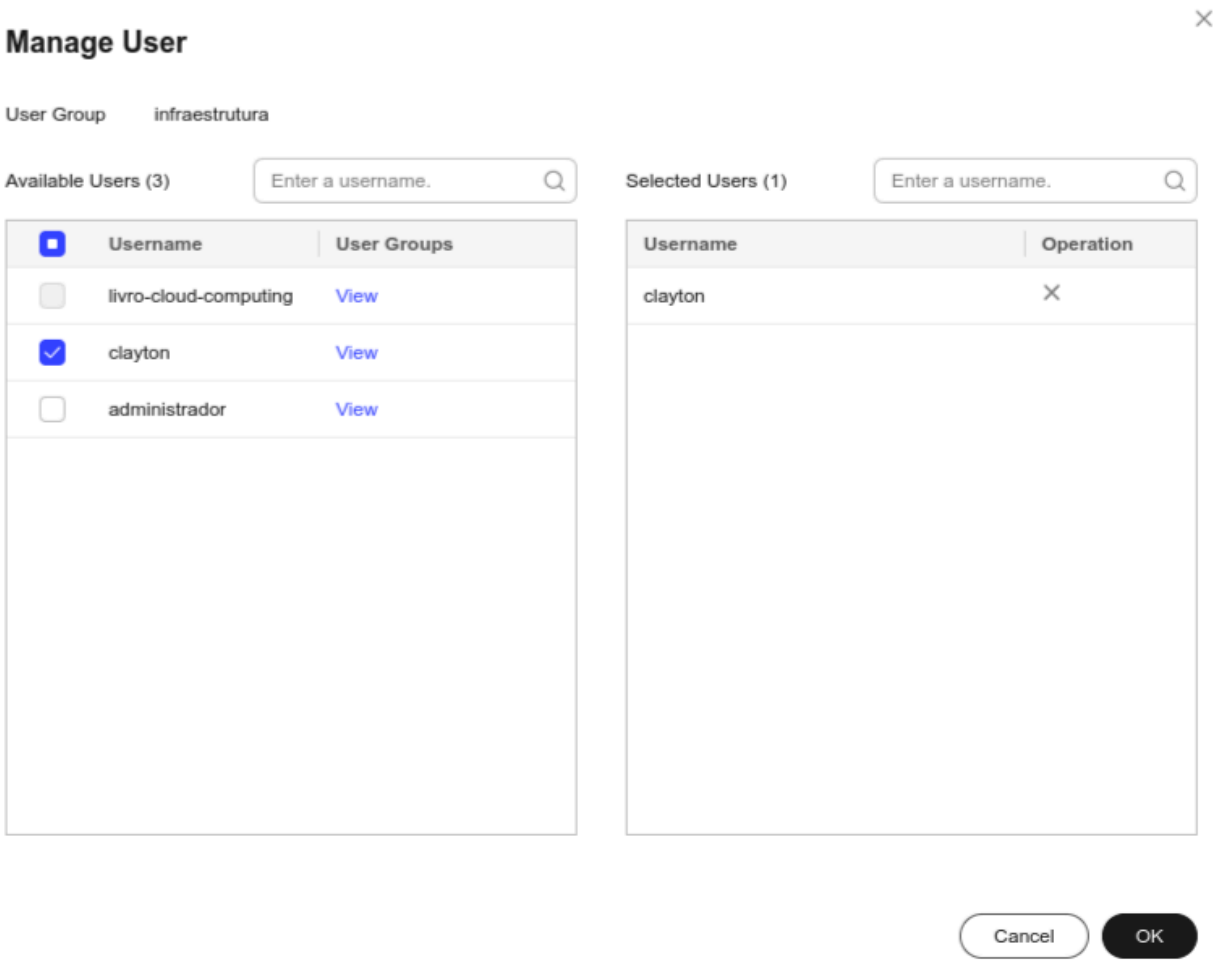


Figura 4.19: Adicionando usuários preexistentes a um novo grupo

Vinculados os usuários desejados ao grupo, precisamos definir quais serão os privilégios recebidos pelos participantes. Essa opção, denominada como definição da política de uso, pode ser acessada por meio da opção "*Authorize*", localizada à direita do grupo que se deseja configurar, na seção "*User Groups*" do *IAM*. De acordo com a interface exibida a seguir, existem cerca de sessenta páginas de políticas predefinidas pelo provedor de nuvem, que permitem a utilização de serviços específicos com diferentes níveis de granularidade de acesso: administrativo, somente leitura, completos, entre outras possibilidades.

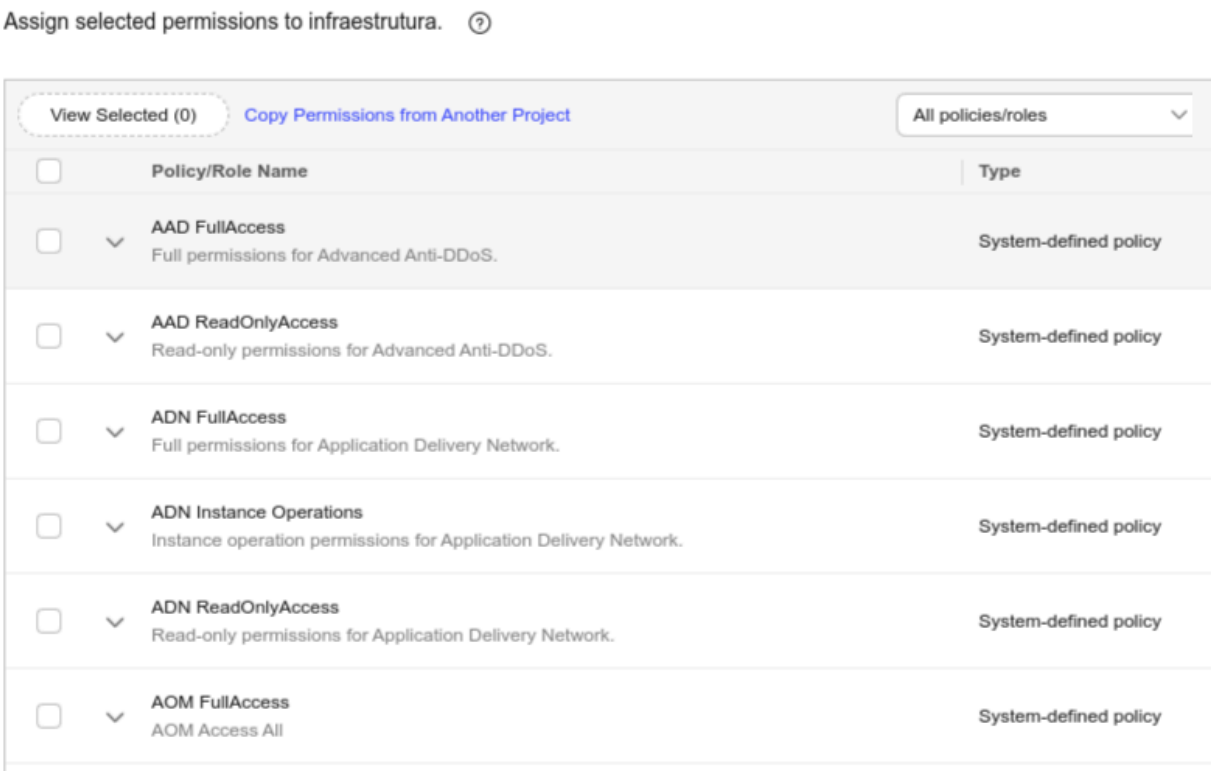


Figura 4.20: Adicionando uma política de uso preexistente ao grupo a ser criado

Vamos imaginar que, a princípio, os usuários pertencentes ao grupo de infraestrutura vão utilizar somente serviços de computação baseados em virtualização para provisionamento de infraestrutura. O serviço responsável por tal funcionalidade na *Huawei Cloud* é o *ECS*. Como queremos garantir a aplicação do princípio do privilégio mínimo, vamos utilizar o termo "*ecs*" como filtro na caixa de buscas, o que possibilita a redução das políticas de acesso preexistentes para somente seis opções. Vamos utilizar o *template* denominado "*ECS FullAccess*", que garante todas as permissões para a operação do serviço em questão.

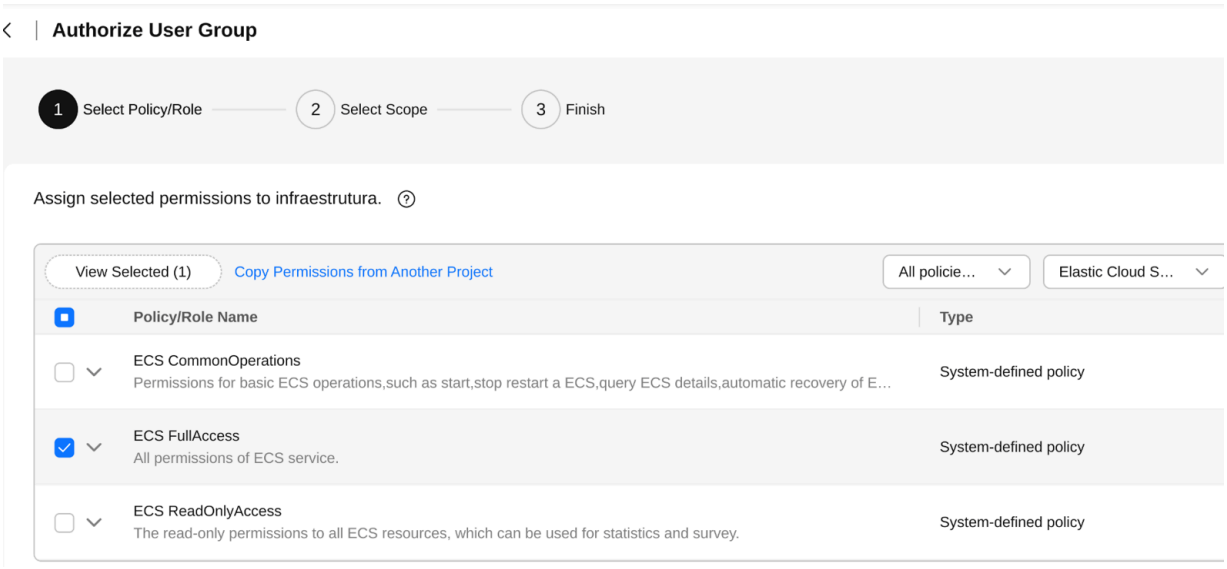


Figura 4.21: Escolhendo uma política aderente ao papel desempenhado pelo grupo

Para que possamos compreender o conceito de granularidade aplicado às políticas de acesso, vamos expandir o *template ECS FullAccess* e visualizar a quantidade de operações que serão permitidas a usuários pertencentes ao grupo em questão. São aproximadamente trinta linhas de especificações de operações.

Authorize User Group

1 Select Policy/Role — 2 Select Scope — 3 Finish

Assign selected permissions to infraestrutura.

View Selected (1) | Copy Permissions from Another Project

Policy/Role Name
ECS CommonOperations Permissions for basic ECS operations,such as start,stop restart a ECS,query ECS
ECS FullAccess All permissions of ECS service.


```
{
    "Version": "1.1",
    "Statement": [
        {
            "Action": [
                "ecs:*:*",
                "evs:*:get",
                "evs:*:list",
                "evs:volumes:create",
                "evs:volumes:delete",
                "evs:volumes:attach",
                "evs:volumes:detach",
                "evs:volumes:manage",
                "evs:volumes:update",
                "evs:volumes:use",
                "evs:volumes:uploadImage",
                "evs:snapshots:create",
                "vpc:*:get",
                "vpc:*:list",
                "vpc:networks:create",
                "vpc:networks:update",
                "vpc:subnets:update",
                "vpc:subnets:create",
                "vpc:ports:*",
                "vpc:routers:get",
```

Figura 4.22: Verificando a granularidade da política escolhida

O próximo passo nos permite escolher qual o alcance ou escopo da política. É possível restringir a autoridade do grupo a projetos locais, limitados a uma ou mais regiões, ou ainda permitir autoridade de acesso a todos os projetos, sem nenhum tipo de restrição.

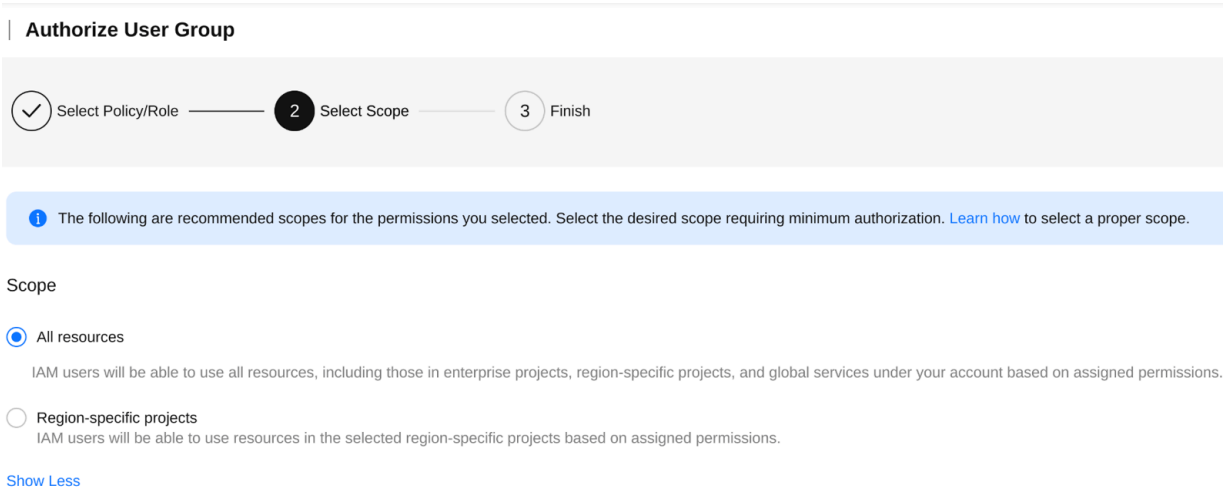


Figura 4.23: Definindo o escopo de abrangência da política

Como de hábito, o último passo apresenta um compilado com as informações pertinentes à operação que acaba de ser concluída.

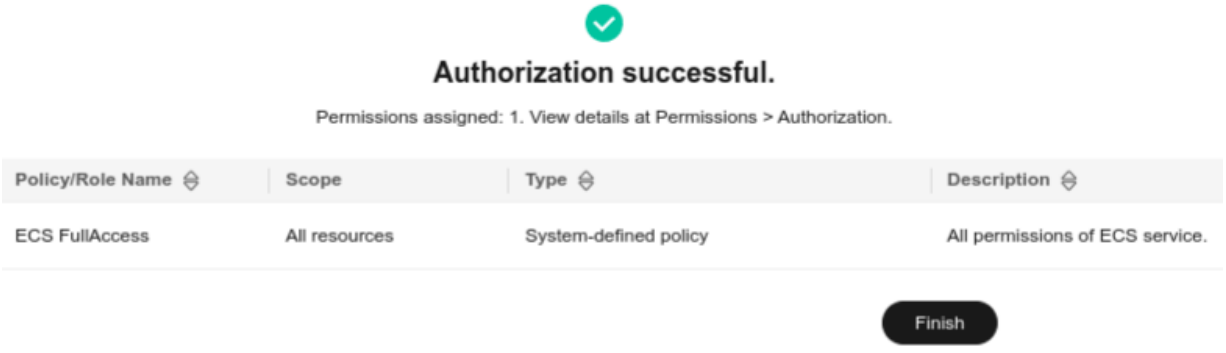


Figura 4.24: Sumário com as propriedades aplicadas

Por fim, ao clicar na opção *Permissions/Authorization* no menu do lado esquerdo, teremos sumarizadas todas as políticas atualmente habilitadas na conta, bem como a quais grupos se referem.

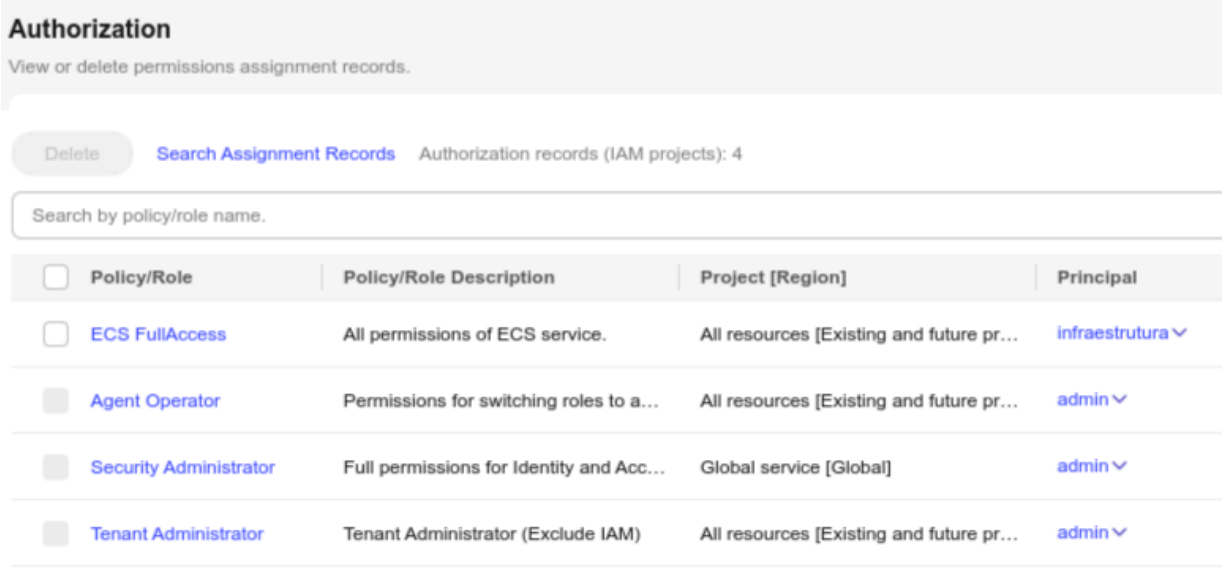


Figura 4.25: Políticas de autorização predefinidas e disponíveis para uso atualizadas

4.4 Considerações importantes

Políticas de acesso que, porventura, estejam em operação e tenham sido aplicadas a grupos no momento em que estes foram criados podem ser modificadas ou removidas a qualquer tempo, de acordo com a dinâmica de uso de recursos do cliente.

Nos exemplos apresentados, os profissionais do grupo de infraestrutura poderiam agregar novas permissões referentes a outros serviços da categoria de computação que, eventualmente, poderiam utilizar *workloads* baseados em contêineres ou computação sem servidor, por exemplo. Do mesmo modo, o usuário "administrador" criado no início do processo com o perfil "*Full permissions*", que concentra os privilégios concedidos pelas políticas "*Agent Operator*", "*Security Administrator*" e "*Tenant Administrator*", simultaneamente, poderia "dividir" essas responsabilidades com outros dois usuários, diminuindo os riscos associados à utilização da conta.

Alguns provedores de nuvem permitem que as políticas de uso sejam aplicadas diretamente ao usuário. Tal prática, apesar de possível, não é recomendada, pois dificulta o processo de gestão e pode induzir a erros ou problemas críticos de segurança, como a não observância do princípio do privilégio mínimo, por exemplo.

Questionário online

Figura 4.26: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 4.27: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo

Parte 3: Serviços básicos

CAPÍTULO 5
Serviços de computação em nuvem

As ferramentas de precificação, autenticação e autorização que tivemos a oportunidade de conhecer nos capítulos anteriores possuem nomenclaturas autoexplicativas e desempenham funções específicas. Em termos práticos, isso significa que, para realizar a precificação de serviços, basta que o usuário utilize uma *Price Calculator*. De mesma maneira, para gerenciar identidades e o acesso a serviços e recursos disponibilizados pelo provedor de nuvem, a utilização do *IAM* é plenamente suficiente.

Os chamados "serviços de computação em nuvem", por outro lado, representam uma categoria de serviços composta por um portfólio diversificado de produtos capazes de executar atividades de processamento — ou computação — especializadas, a fim de atender a diferentes fluxos de trabalho e necessidades de negócio específicas. Nesse sentido, é fundamental que o usuário entenda que os diferentes serviços ofertados dentro da mesma categoria não competem entre si, mas, sim, se complementam.

Qualquer provedor de nuvem pública e toda solução de *software* voltada à implementação de nuvens privadas certamente vão contar com pelo menos três serviços de computação voltados a fluxos de trabalho específicos, a saber:

- Serviços de computação em nuvem baseados em contêineres;
- Serviços de computação em nuvem baseados em eventos;
- Serviços de computação em nuvem baseados em virtualização.

Em geral, as tecnologias utilizadas para provisionamento de infraestrutura em ambientes de nuvem são as mesmas utilizadas no modelo de infraestrutura própria. No entanto, ao optar pelo modelo de nuvem, o usuário não precisa lidar com a complexidade envolvida nos processos de gestão da infraestrutura física dos equipamentos, uma vez que se trata de uma responsabilidade atribuída ao provedor de nuvem.

De mesma maneira, ambientes complexos de computação distribuída — que contam com arquiteturas baseadas em *clusters*, voltadas para orquestração de contêineres, o balanceamento de carga e a alta disponibilidade, por exemplo — podem ser utilizados sem grande complexidade de implantação, graças às interfaces intuitivas do console de gerenciamento *web*.

Partindo dessa lógica, o cliente deve ter clareza sobre quais tipos de arquiteturas e serviços deseja utilizar em suas novas aplicações, antes do efetivo provisionamento de infraestrutura em ambiente de nuvem. A depender das tecnologias escolhidas, torna-se necessária a utilização de um ou de outro serviço de computação específico, disponibilizado pelo provedor.

Quando se trata de migração de aplicações legadas, é possível que o cliente opte por realizar uma migração *lift and shift*, na qual a infraestrutura original é migrada rapidamente para o ambiente de nuvem, pois não requer alterações significativas. Alternativamente, o cliente pode aproveitar a oportunidade para refatorar a aplicação, utilizando um modelo de desenvolvimento mais moderno, que aproveite melhor o paradigma de computação em nuvem.

Lembre-se: a escolha de tecnologias assertivas, que proporcionem o desenvolvimento de aplicações *cloud native*, é muito importante para que os resultados esperados em um ambiente de computação em nuvem sejam plenamente alcançados.

5.1 Serviços de computação em nuvem baseados em virtualização

Os serviços de computação baseados em virtualização são os mais populares dentro da categoria, pois, além de servirem como base para outros serviços agregados, representam, na maioria dos casos, o primeiro contato de um novo usuário com um serviço disponibilizado pelo provedor de computação em nuvem.

A literatura é repleta de referências que evidenciam os benefícios da virtualização frente às alternativas tradicionais de implantação de servidores dedicados. Esse recurso torna-se ainda mais interessante em ambientes de nuvem, pois a utilização do modelo de infraestrutura como serviço libera o cliente da complexidade envolvida na gestão das ferramentas que tornam o uso da tecnologia possível.

O serviço de virtualização disponibilizado pela *Huawei Cloud* recebe o nome de "*Elastic Cloud Server*" (ou simplesmente *ECS*). Assim como outros serviços de mesma natureza, essa ferramenta requer que o usuário informe quais os parâmetros necessários para a criação de uma nova instância de virtualização.

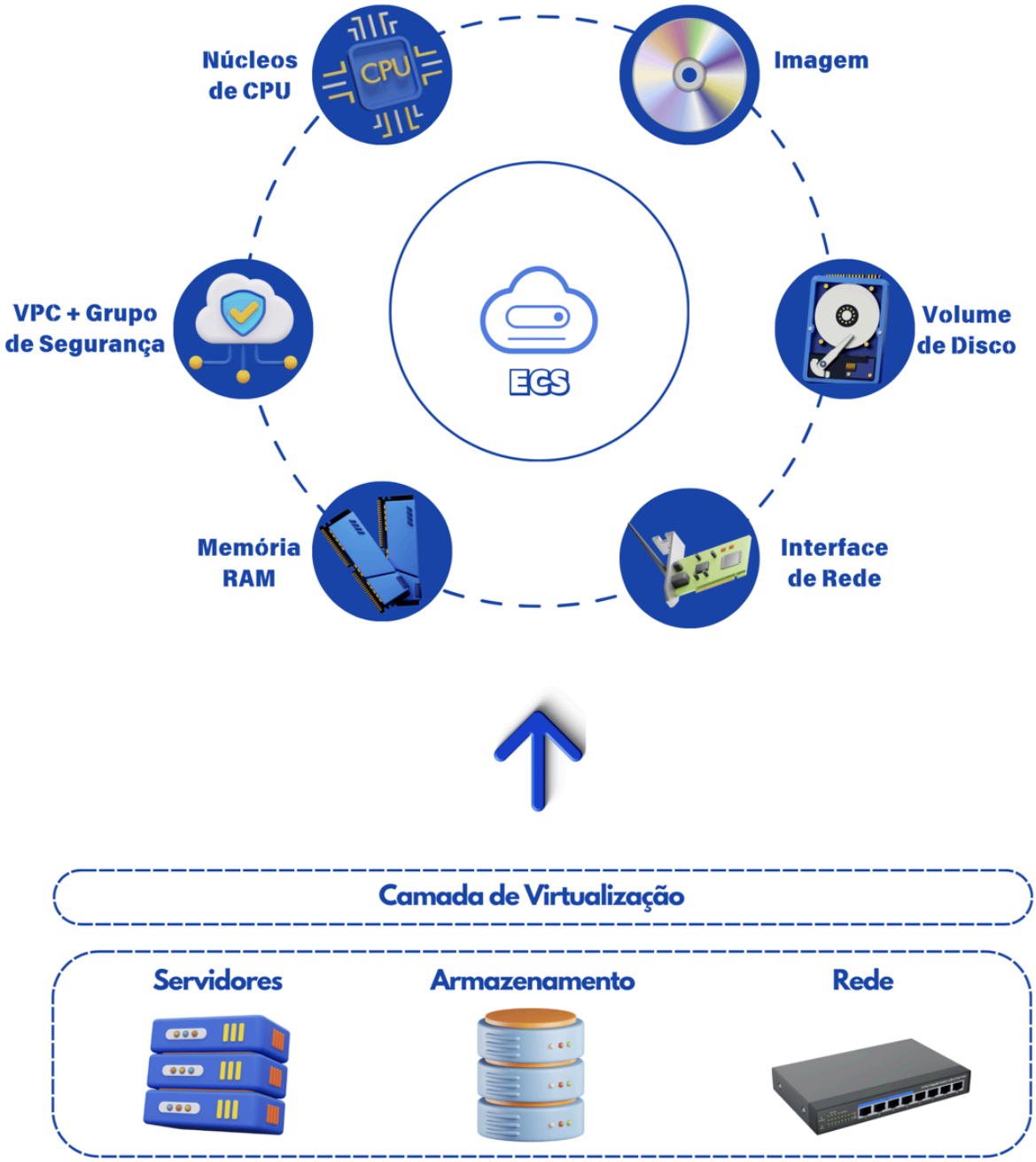


Figura 5.1: Elastic Cloud Server - ECS, da Huawei Cloud

Conforme mencionado anteriormente, todo provedor de nuvem pública ou solução de nuvem privada certamente ofertará a seus usuários um serviço de computação baseado em virtualização. Em geral, a utilização desse tipo de solução não tem grandes mistérios. Por trás das cortinas, empregam-se as mesmas soluções de

virtualização profissional utilizadas no modelo de infraestrutura própria. No entanto, no ambiente de nuvem, o usuário conta com novas interfaces de gerenciamento que tornam o processo de criação de instâncias de virtualização mais intuitivo e versátil.

A interface de gerenciamento *web* permite que o usuário informe os parâmetros necessários para a criação de uma nova instância, utilizando componentes gráficos tipicamente existentes em interfaces equivalentes. Já os acessos programáticos podem ser realizados por meio de utilitários de linha de comando, *APIs* e *SDKs* específicos, disponibilizados pelos provedores.

Criando uma instância de virtualização

Neste tópico, vamos explicar detalhadamente o processo de criação de uma instância de virtualização. Para tanto, utilizaremos como referência o serviço de nuvem pública disponibilizado pela *Huawei Cloud*, o *ECS*.

Para visualizarmos os serviços oferecidos pelo provedor de nuvem, vamos utilizar o console de gerenciamento *web*, clicando no ícone disponível no canto superior esquerdo da tela, composto por três barras horizontais. Essa ação expandirá a lista de categorias de serviços disponíveis, bem como as opções presentes em cada uma delas. No nosso caso, vamos utilizar o serviço de computação denominado "*Elastic Cloud Server*", disponível na seção "*Compute*".

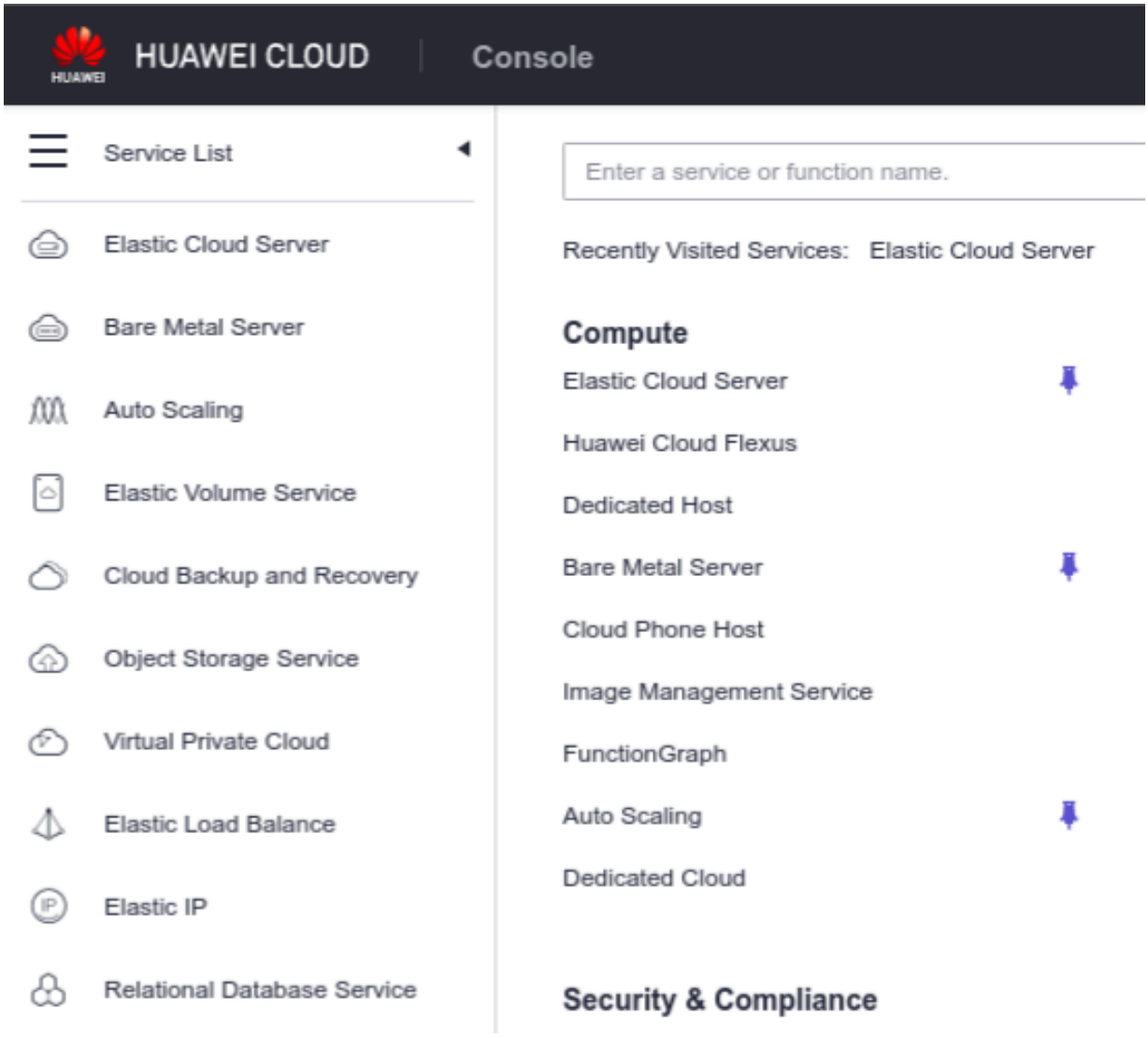


Figura 5.2: Console de Gerenciamento Web

Ao selecionarmos o serviço em questão, é apresentada uma interface inicial com informações sobre as instâncias eventualmente preexistentes na região em que o usuário está logado. Os dados disponibilizados nessa seção são organizados em formato de tabela e reúnem os parâmetros operacionais mais relevantes referentes a cada uma das instâncias. Como estamos lidando com o serviço pela primeira vez, obviamente, não existem dados prévios.

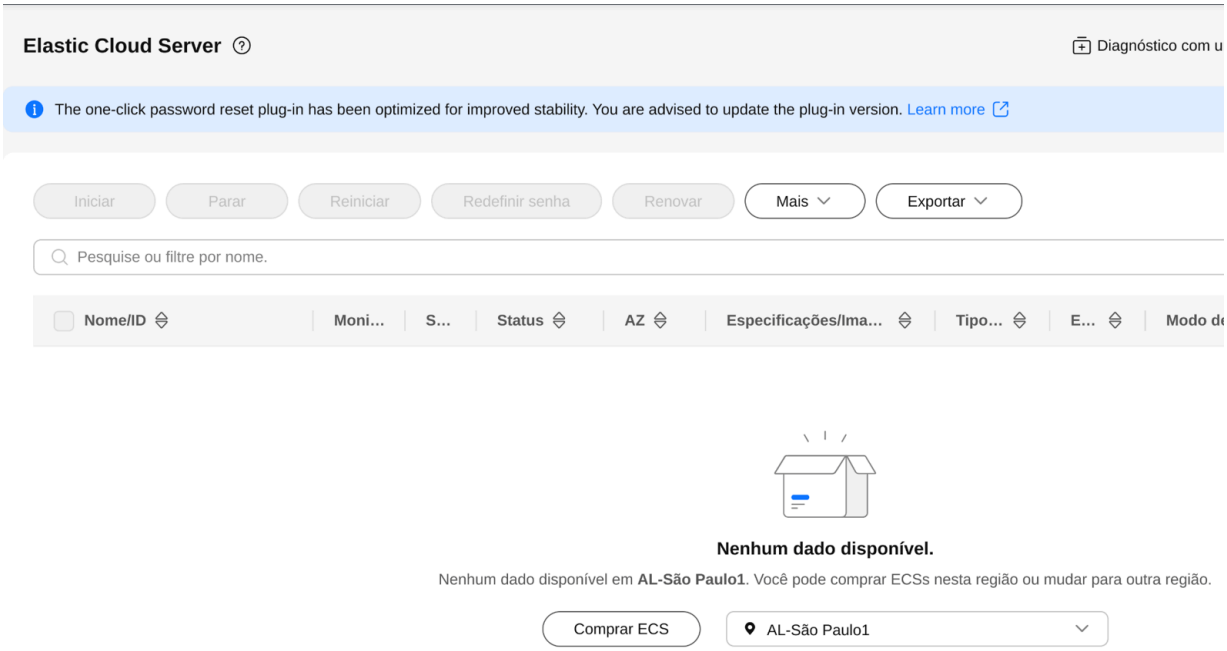


Figura 5.3: Elastic Cloud Server - ECS, da Huawei Cloud

O *dashboard*, como o próprio nome sugere, é um painel mais informativo, que apresenta informações de forma mais dinâmica e intuitiva à medida que novos recursos de infraestrutura são criados. Ao clicar no botão "*Buy ECS*", iniciaremos o processo de criação de uma nova máquina virtual que, por convenção, passaremos a chamar de "instância de virtualização".

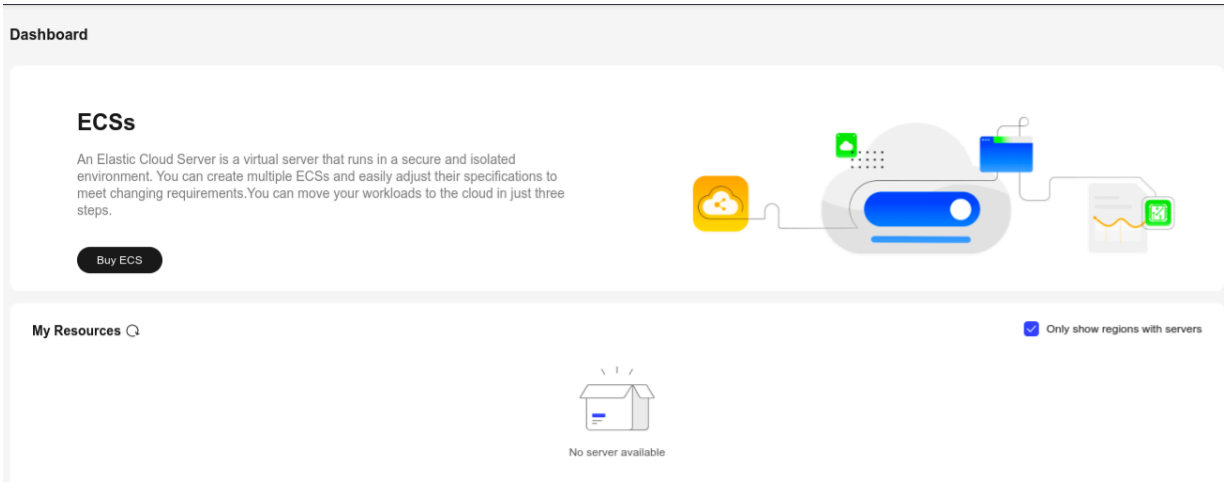


Figura 5.4: Dashboard do ECS

Novos usuários devem ter em mente que, seja em um ambiente de nuvem pública ou ao optar por ferramentas de gestão de uma nuvem privada, o objetivo é sempre tornar a vida do usuário mais simples.

A oferta de assistentes intuitivos tende a ser o padrão nesse paradigma. Eles solicitam somente os parâmetros estritamente necessários para o provisionamento do serviço que se deseja implementar, utilizando o clássico modelo "*next, next, finish*", comum em ferramentas *desktop* voltadas ao usuário final.

Nesse sentido, revisaremos alguns parâmetros que, embora já tenham sido apresentados em capítulos anteriores, agora serão alinhados com o propósito da ferramenta em questão. Outros parâmetros serão abordados apenas de forma superficial neste momento, pois serão aprofundados oportunamente, em capítulos posteriores.

O processo de criação de uma instância de virtualização exige, inicialmente, que o usuário defina a região em que o recurso será criado. Essa decisão depende de vários fatores, como precificação, latência, disponibilidade de serviços, conformidades, entre outras questões complexas que merecem reflexão. Como o objetivo aqui é somente ilustrar o processo de criação de uma nova instância, não nos aprofundaremos nesse aspecto. No entanto, é importante ter em mente que, em um cenário real, a escolha adequada de cada parâmetro é fundamental para o sucesso e a continuidade de um projeto.

Dado o contexto, naturalmente vamos escolher a região mais próxima — no caso, "*Latin America — São Paulo 1*". Em seguida, definiremos qual o método de pagamento mais conveniente às nossas necessidades. Lembre-se de que as interfaces dos serviços normalmente disponibilizam ícones do tipo "*hint*", que trazem informações valiosas acerca do item que estamos prestes a escolher.

Nesse caso específico, as opções mais comuns permitem que selecionemos o pagamento conforme o uso ("*pay-per-use*") ou, ainda, que façamos o pagamento com periodicidade mensal ou anual, adotando a modalidade de aquisição de instâncias reservadas.

Os conceitos envolvidos, vantagens e desvantagens, bem como os cenários de uso recomendados para cada modalidade, foram amplamente discutidos no capítulo sobre precificação, no qual tivemos a oportunidade de simular uma série de situações.

Em uma aplicação real, é comum validar os parâmetros de uma instância por um determinado período, para que, uma vez devidamente testados e homologados, eles passem a adotar um modelo de precificação mais vantajoso.

A regra de ouro aqui consiste em revisitar as instâncias periodicamente, com vistas ao monitoramento constante dos pilares de excelência operacional e de otimização de custos. Evidentemente, à medida que sua infraestrutura aumenta, automatizar esse tipo de verificação torna-se algo mandatório. Veremos como realizar esse processo em capítulos futuros.

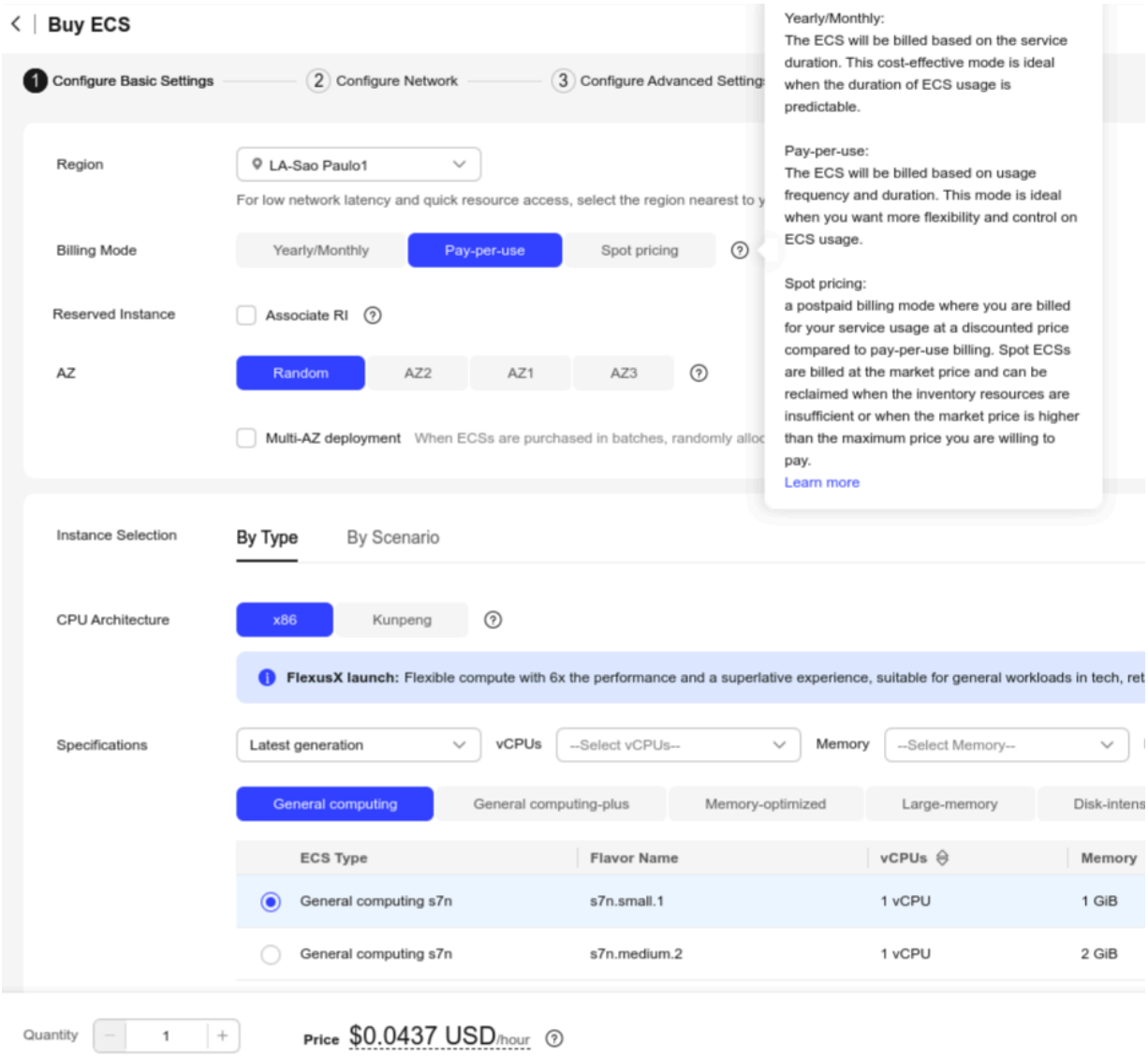


Figura 5.5: Billing Modes disponíveis no ECS

A definição da zona de disponibilidade utilizada dentro da região pode ser feita de forma randômica ou pontual, conforme a necessidade ou preferência do usuário. Vale lembrar que o provisionamento de infraestrutura no modelo "*Multi-AZ*" — ou seja, com réplicas idênticas em diferentes zonas de disponibilidade — proporciona alta disponibilidade e, consequentemente, maior tolerância a falhas.

Trata-se, portanto, de uma prática altamente recomendada. Nunca é demais lembrar que replicar infraestrutura em múltiplas zonas de disponibilidade ou regiões aumenta os custos envolvidos na mesma proporção. Tenha sempre isso em mente ao precificar seu projeto e tome cuidado para não abrir mão de recursos de segurança fundamentais.

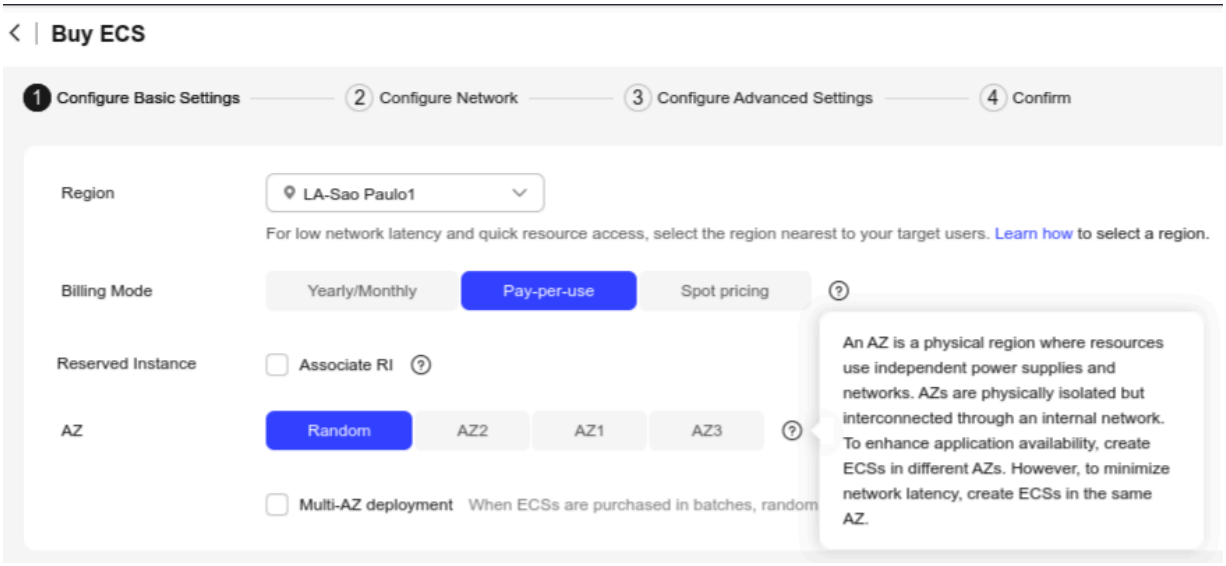


Figura 5.6: Definição das zonas de disponibilidade

Conforme vimos anteriormente, os provedores de nuvem tendem a disponibilizar ao cliente uma série de perfis de configuração de instância predefinidos, alinhados ao propósito da infraestrutura a ser criada. Esses perfis variam desde máquinas destinadas à computação de propósito geral até outras otimizadas para ampla utilização de memória, armazenamento em disco, computação de alto desempenho, entre várias opções disponíveis.

Dentro de cada grupo, há ainda variações com diferentes gerações e arquiteturas de processadores, quantidade de *cores*, quantidade de memória *RAM*, performance de rede, entre outros parâmetros que, obviamente, vão influenciar no custo da instância.

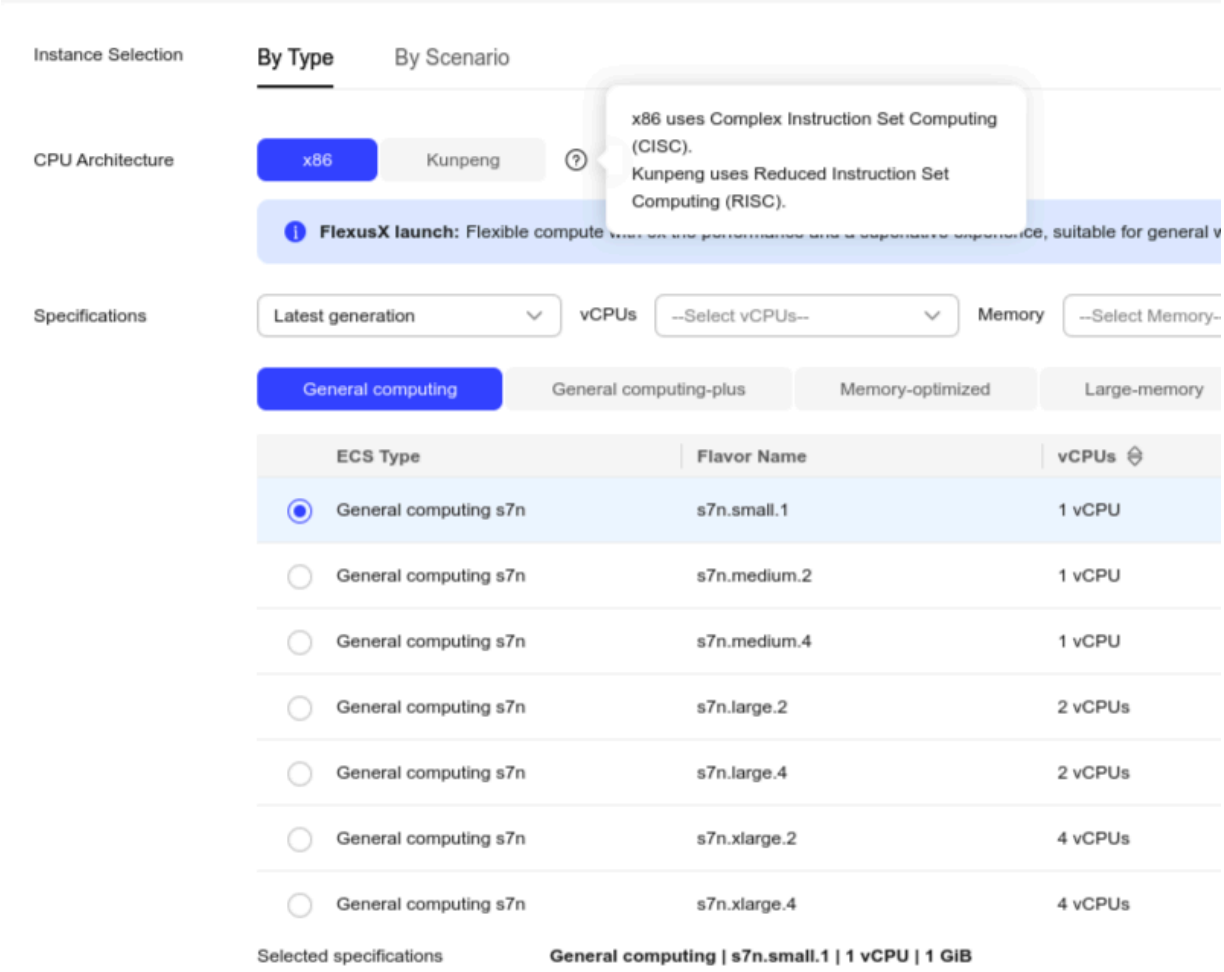


Figura 5.7: Escolha do tipo de instância de acordo com o tipo

Da mesma maneira, os perfis podem ser organizados de acordo com o tipo específico de aplicação, descritos de forma mais específica — como servidores *web*, bancos de dados, análise de dados, inteligência artificial, renderização de imagens, entre outros cenários de aplicação. Evidentemente, o objetivo da disponibilidade de tais perfis é apenas nortear as escolhas do cliente, ofertando configurações compatíveis com cada cenário de aplicação. Como não estamos lidando com uma ciência exata, cabe ao usuário o cuidado e a responsabilidade de monitorar frequentemente os indicadores de utilização de suas instâncias, a fim de verificar se os perfis predefinidos, de fato, atendem às suas necessidades específicas.

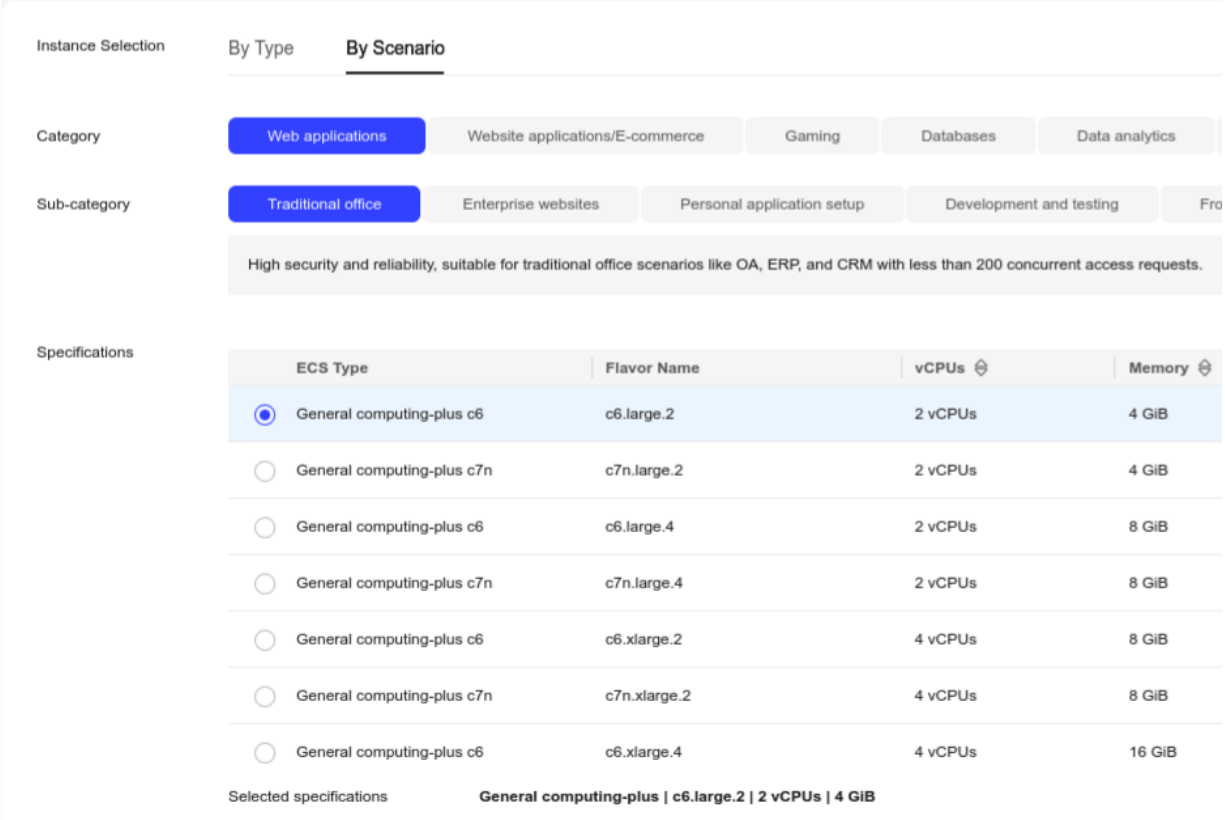


Figura 5.8: Escolha do tipo de instância de acordo com o cenário de aplicação

É possível ainda filtrar os *templates* de configuração disponíveis de acordo com os requisitos desejados, sendo possível escolher opções referentes à arquitetura do processador, quantidade de *cores* e de memória *RAM*.

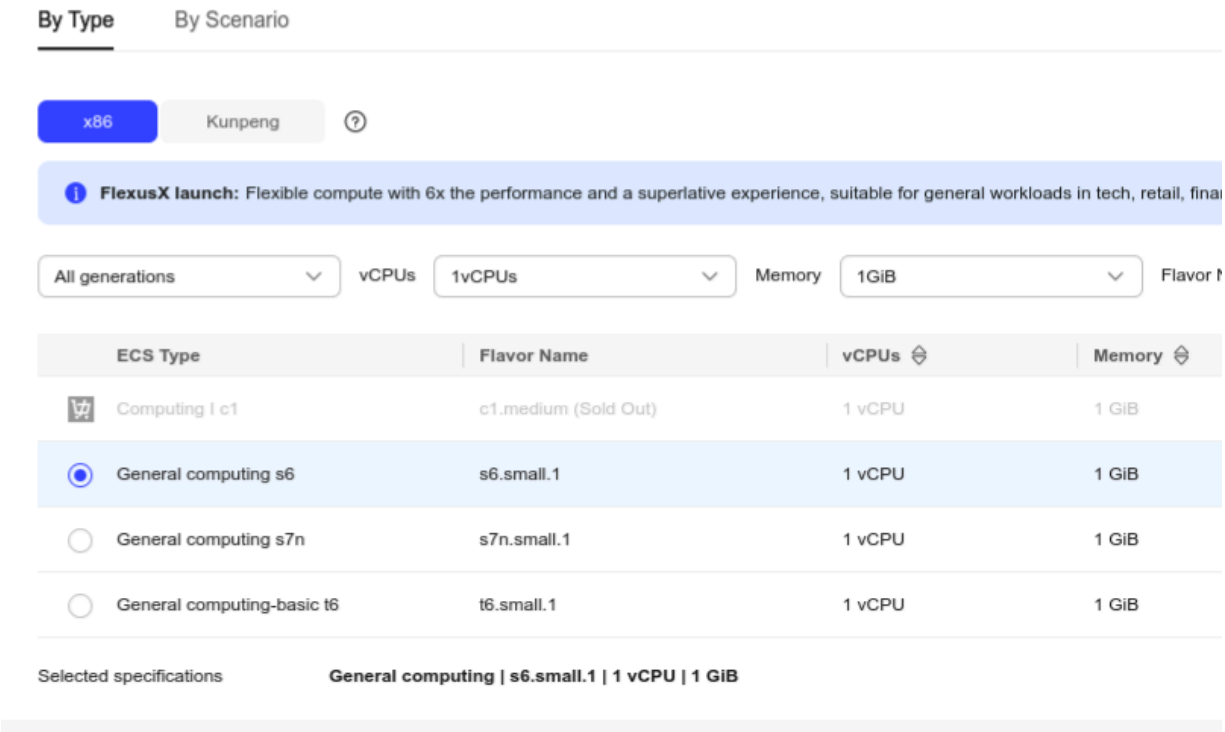


Figura 5.9: Escolha do tipo de instância de acordo com os parâmetros

Definidos os parâmetros de *hardware* desejados, é hora de especificar qual sistema operacional será utilizado na instância. Trata-se de uma imagem auditada e disponibilizada pelo provedor de nuvem na seção de imagens públicas. Tais imagens não requerem nenhum tipo de interação por parte do usuário para instalação, pois são baseadas em *snapshots* funcionais. Eventuais custos de licenciamento do sistema escolhido já estão incorporados à modalidade de pagamento escolhida pelo usuário.

Em um computador pessoal, o usuário precisa instalar o sistema operacional de sua preferência por meio de um processo assistido, informando parâmetros do equipamento a ser utilizado, as eventuais ferramentas que deseja usar, bem como a localização dos arquivos no disco rígido, por exemplo.

Em ambiente de nuvem, o provedor utiliza ferramentas que permitem reutilizar um sistema operacional recém-instalado, funcional e pronto para uso, sem a necessidade de repetir o processo de instalação assistido. Para tanto, é gerada uma "fotografia" — *snapshot* — do sistema de arquivos, que reflete um ponto de recuperação específico.

As distribuições de *snapshots* homologadas pelo provedor, que podem ser restauradas em volumes de armazenamento de *boot*, recebem o nome de imagens. Elas possibilitam que instâncias de virtualização estejam disponíveis para utilização em questão de minutos.

Vale lembrar que o cliente pode criar suas próprias imagens, derivadas das imagens públicas disponibilizadas pelo provedor, promovendo alterações e personalizações que desejar para reutilização e uso futuro. As imagens criadas podem ser mantidas para uso exclusivo do cliente, podendo ser recuperadas na seção "*private images*". Caso o cliente deseje compartilhar a imagem com outros usuários, poderá fazê-lo definindo-a como uma "*shared image*". É importante lembrar que as únicas imagens auditadas e, portanto, recomendadas pelo provedor são as disponibilizadas na seção "*public images*", sendo a utilização das demais de total responsabilidade do cliente.

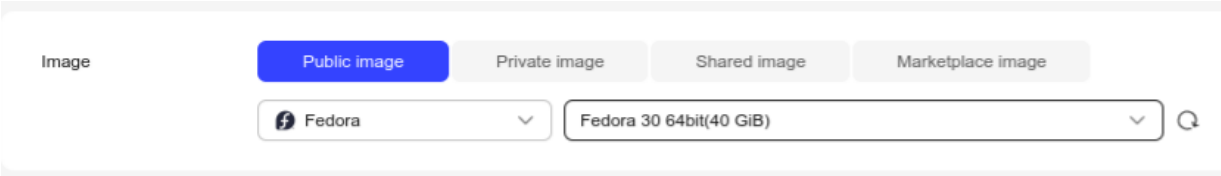


Figura 5.10: Definição da imagem de sistema operacional utilizada

No processo de migração de um *datacenter* próprio para um novo cenário baseado no provisionamento de infraestrutura em ambiente de nuvem pública, é muito comum que usuários se sintam

desmotivados em um primeiro momento, devido à suposta necessidade de abrir mão de soluções de *software* ou *hardware* utilizadas com sucesso por anos em suas organizações. Evidentemente, essa também é uma preocupação dos fabricantes que ofereciam as referidas soluções em ambientes de infraestrutura própria. Para contornar essa questão, os provedores de nuvem pública disponibilizam aos fabricantes um *marketplace*, que permite aos desenvolvedores oferecer suas soluções em formato de *appliances*, possibilitando aos clientes a continuidade de utilização de produtos e soluções de forma virtualizada, em ambiente de nuvem, com precificação embutida na imagem utilizada.

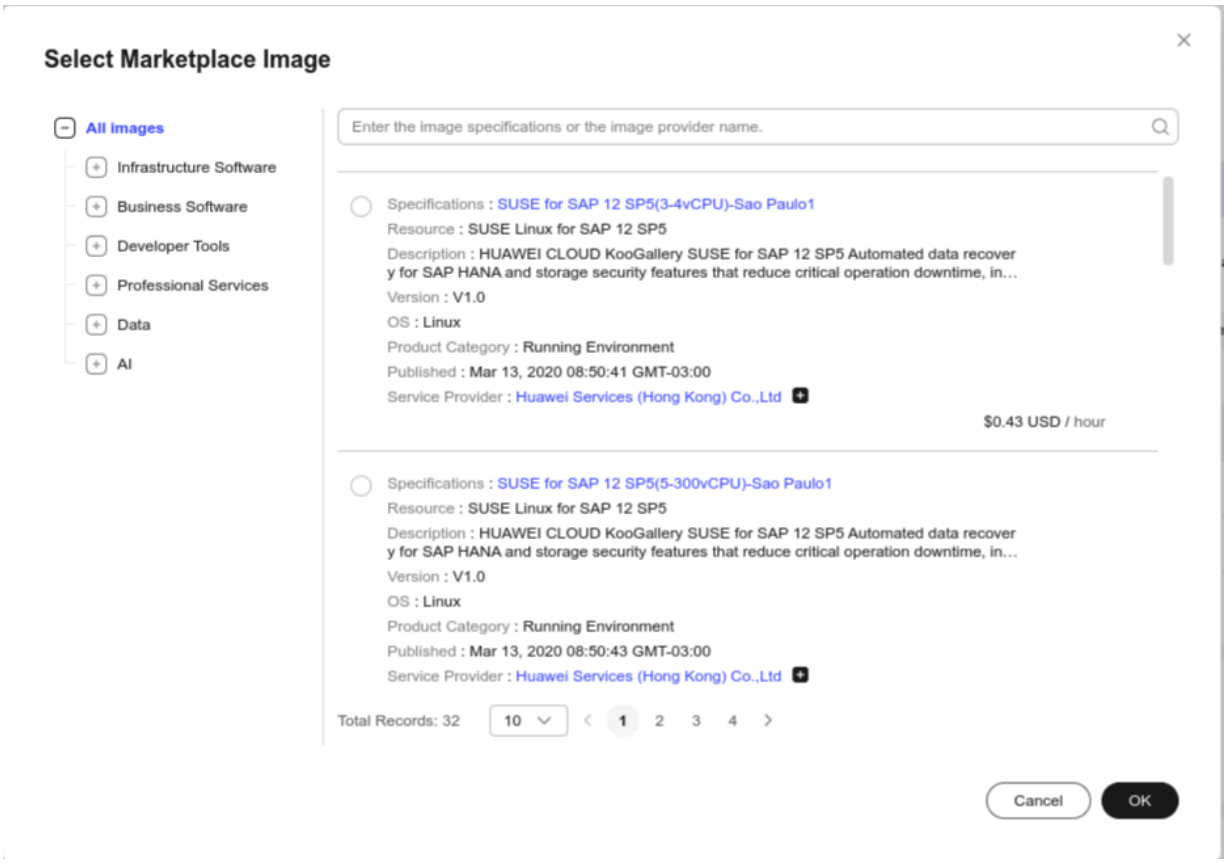


Figura 5.11: Utilizando imagens disponíveis no marketplace

Na sequência, vamos determinar o espaço de endereçamento no qual a instância de virtualização será provisionada. Para tanto, devemos escolher uma *Virtual Private Cloud* dentro da região atual,

bem como uma sub-rede pertencente ao prefixo da *VPC* escolhida. Em resumo, essa etapa é responsável por determinar os parâmetros de rede necessários para a utilização da instância de virtualização.

Da mesma maneira, para que os serviços eventualmente implementados pelo usuário funcionem como esperado na instância que estamos configurando, devemos determinar quais portas deverão estar liberadas para conexões entrantes. Isso porque, pelo princípio do privilégio mínimo, toda nova instância utiliza uma política de acesso fechada, permitindo apenas o tráfego de dados explicitamente autorizado. Os serviços que acabamos de mencionar, assim como outros relacionados ao tema, serão discutidos em detalhes no capítulo de "Serviços de rede".

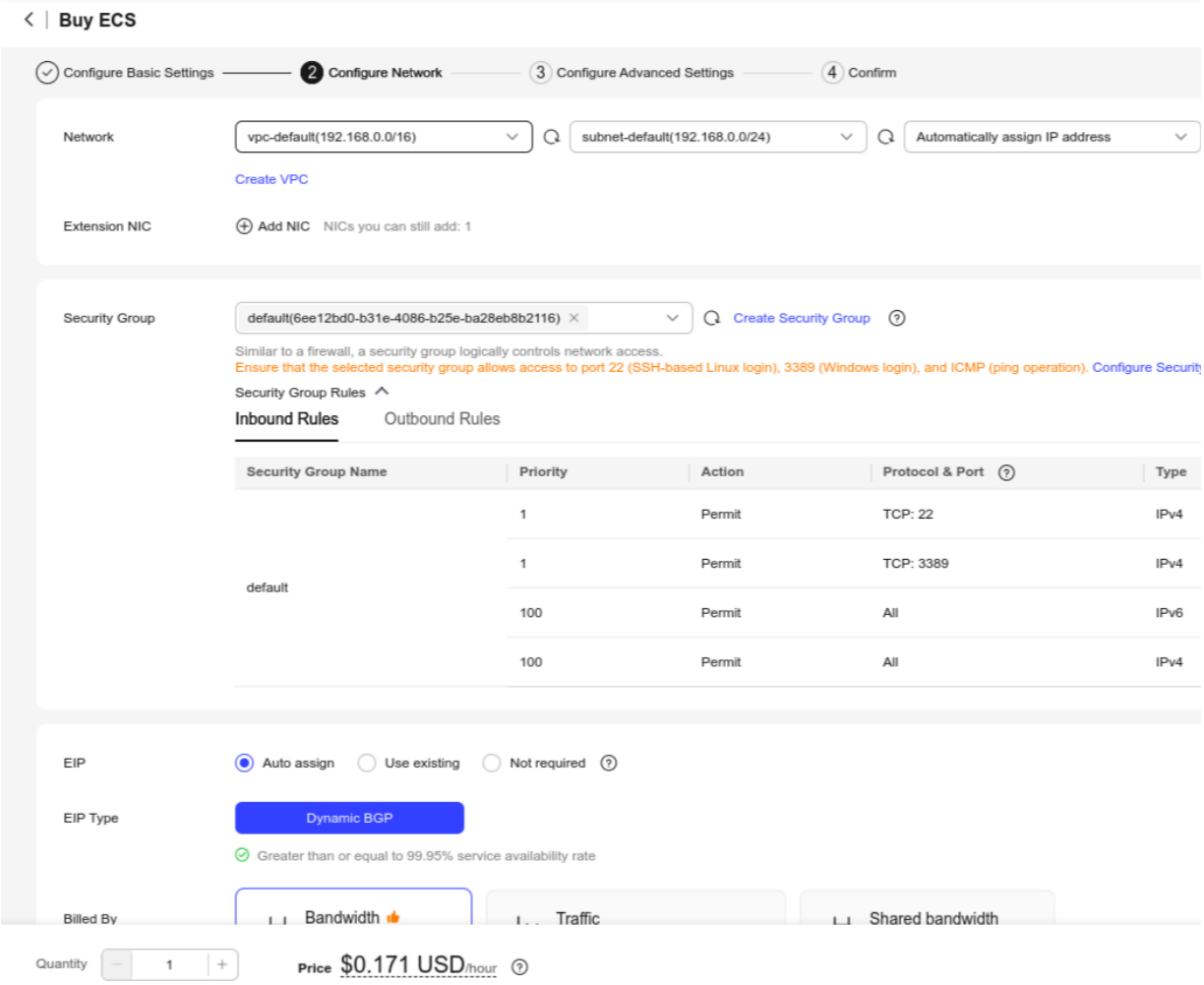


Figura 5.12: Definição dos parâmetros de rede

Um endereço *IP* elástico, ou simplesmente *EIP*, associa um *IP* fixo a uma instância de virtualização. Esse recurso não é obrigatório, mas, uma vez selecionado, gerará tarifas relativas à transferência de dados. Uma das modalidades de contratação baseia-se na utilização de largura de banda por hora. Nessa modalidade, o usuário deve determinar qual o tamanho da largura de banda desejada, e o custo de transferência de dados já estará automaticamente incluído no valor por hora associado à instância, sem custos adicionais.

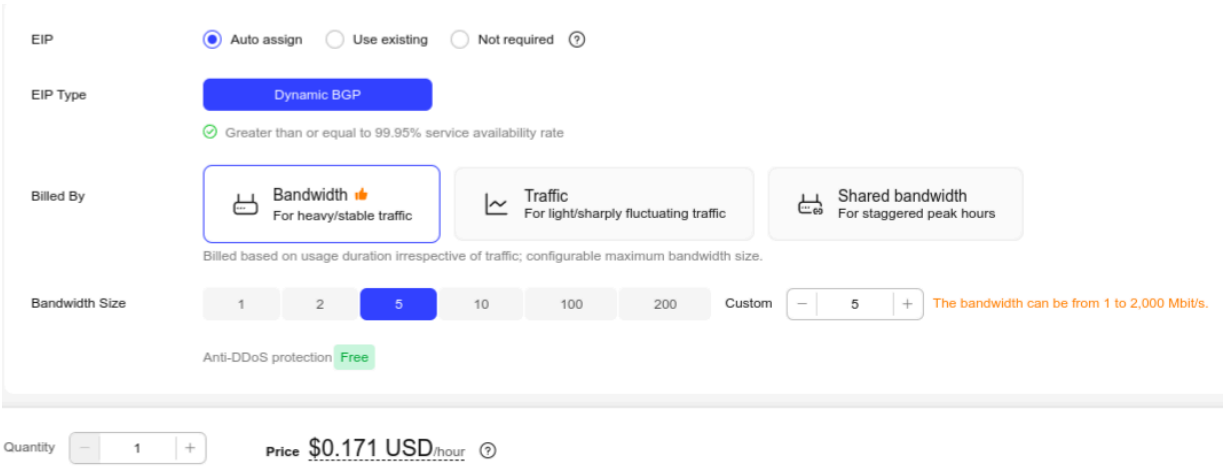


Figura 5.13: Definição do IP elástico de acordo com a largura de banda

A transferência de dados também pode ser tarifada com base na quantidade de dados transferidos, de modo que o usuário será cobrado de acordo com o tráfego de saída total, independentemente do tempo de utilização. Nesse caso, é necessário especificar a largura de banda máxima desejada. Vale lembrar que essa é uma modalidade de cobrança adicional e não está incluída no valor por hora exibido no rodapé da página ou no sumário apresentado ao usuário no momento de criação efetiva da instância, tendo em vista que será cobrada sob demanda, mês a mês. Leve isso em conta ao utilizar essa modalidade.

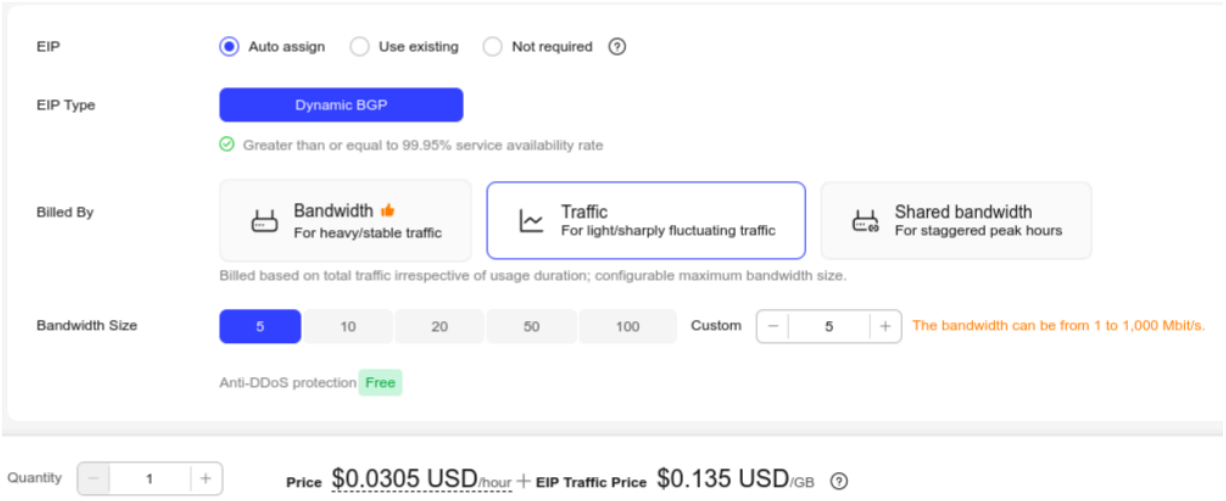


Figura 5.14: Definição do IP elástico de acordo com o tráfego

Na terceira etapa de configuração, vamos inicialmente definir um nome e uma descrição para a instância. Considera-se uma boa prática seguir uma estratégia predefinida para a escolha da nomenclatura associada a cada um dos "objetos de nuvem" presentes em sua infraestrutura. Em geral, recomenda-se utilizar nomes significativos, que façam sentido conforme o serviço ou o papel desempenhado pelo recurso a ser criado, ou seguir um padrão de nomenclatura interno definido pela política de uso de recursos de tecnologia da informação da empresa, por exemplo. A descrição segue a mesma lógica, pois, à medida que a infraestrutura aumenta, mais útil se tornam as informações referentes a nossos ativos em funcionamento.

No momento em que a instância for iniciada e estiver pronta para utilização, é preciso definir qual o método de autenticação desejado para realizar a administração remota do recurso. A grande maioria dos provedores de nuvem pública adota como padrão o uso de um par de chaves assimétricas — uma chave pública e uma privada — para o processo de autenticação, desabilitando o tradicional processo baseado em usuário e senha. Nesse caso, a chave pública fica armazenada na instância, enquanto a privada fica em poder do usuário. O par de chaves pode ser criado juntamente da instância ou reutilizado, caso o usuário já possua um par de chaves que deseje reutilizar. Essa metodologia é recomendada porque elimina ataques de força bruta, nos quais mecanismos automatizados de ataque tentam adivinhar usuário e senha para obter acesso não autorizado à instância.

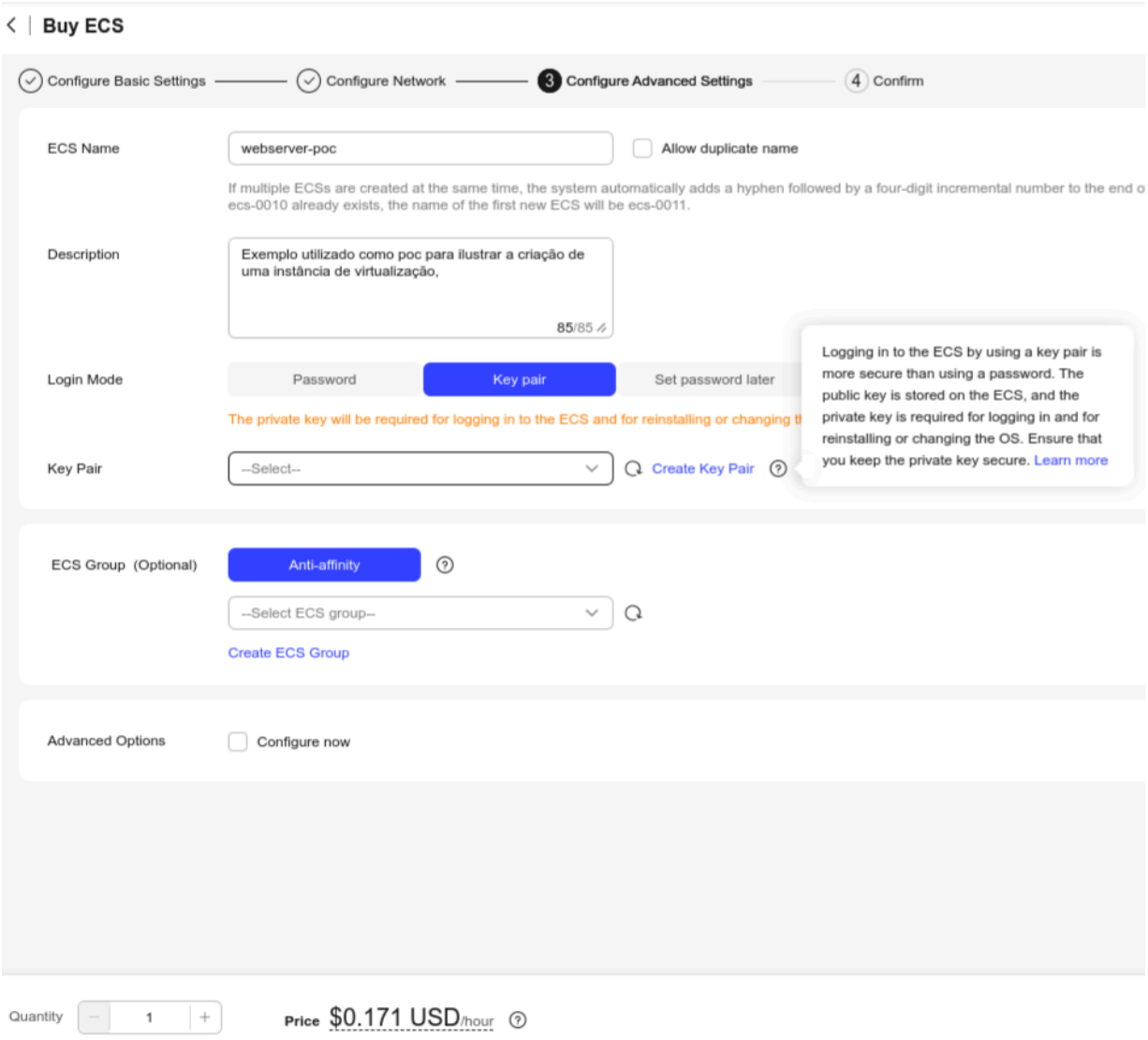


Figura 5.15: Personalizando configurações da instância

Um outro fator agravante, presente na utilização de senhas para autenticação remota, é a eventual utilização do usuário *root* no sistema operacional da instância. Caso o atacante descubra a senha desse usuário, terá plenos poderes no sistema operacional do *host* alvo.

Vale lembrar que, uma vez definido o usuário *root* para fins administrativos, o atacante nem precisa descobrir o nome do usuário em uso, pois ele já é predefinido pelo provedor de nuvem, restando apenas descobrir a senha para acesso. Ainda que a senha utilizada

deva atender a requisitos predefinidos na política de utilização de senhas dos usuários, essa modalidade de acesso não é considerada uma boa prática de segurança.

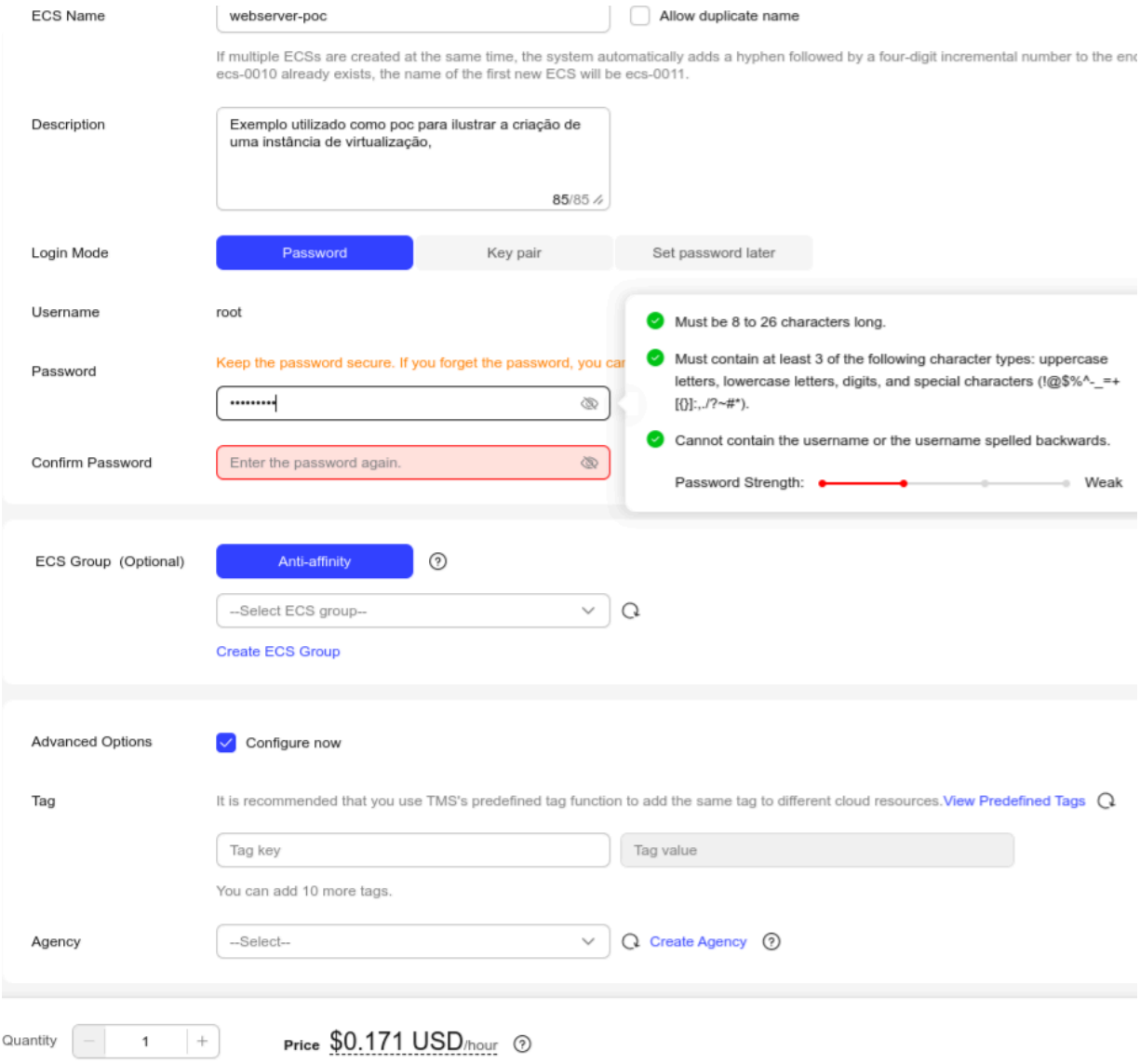


Figura 5.16: Configurando a senha do usuário root

As *tags* são um recurso valioso para a administração de recursos provisionados em ambiente de nuvem. Sua utilização está associada à atribuição de "etiquetas" lógicas, baseadas em chave-valor, aos objetos de nuvem que se deseja identificar. Essa simples ação permite que os ativos associados às *tags* sejam facilmente

localizados por meio de operações de filtragem ou que operações de manutenção sejam feitas em lote, utilizando as chaves desejadas como fator de seleção, desde que, obviamente, o provedor de nuvem disponibilize esse tipo de ferramenta ao administrador.

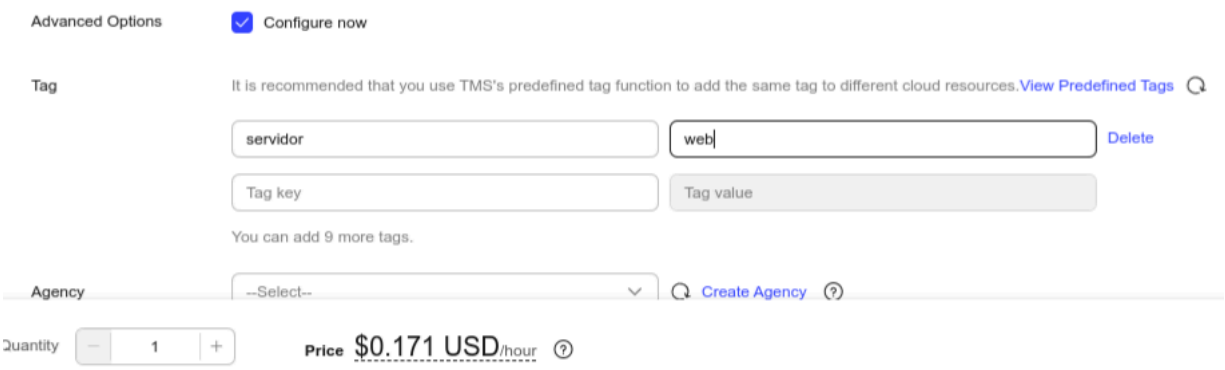


Figura 5.17: Definindo tags chave-valor

Em ambientes corporativos, não são raras as situações em que usuários que não fazem parte do quadro de colaboradores, como auditores, por exemplo, tenham acesso concedido a equipamentos e sistemas em caráter temporário. Nesse tipo de cenário, considera-se uma boa prática adotar mecanismos de autenticação que mantenham o acesso por um período de tempo predeterminado, de modo que o operador em questão não seja cadastrado como um usuário regular. Alguns provedores de nuvem utilizam entidades específicas para viabilizar esse tipo de funcionalidade, como uma "*agency*", no caso da *Huawei Cloud*, ou "funções", em outros provedores.

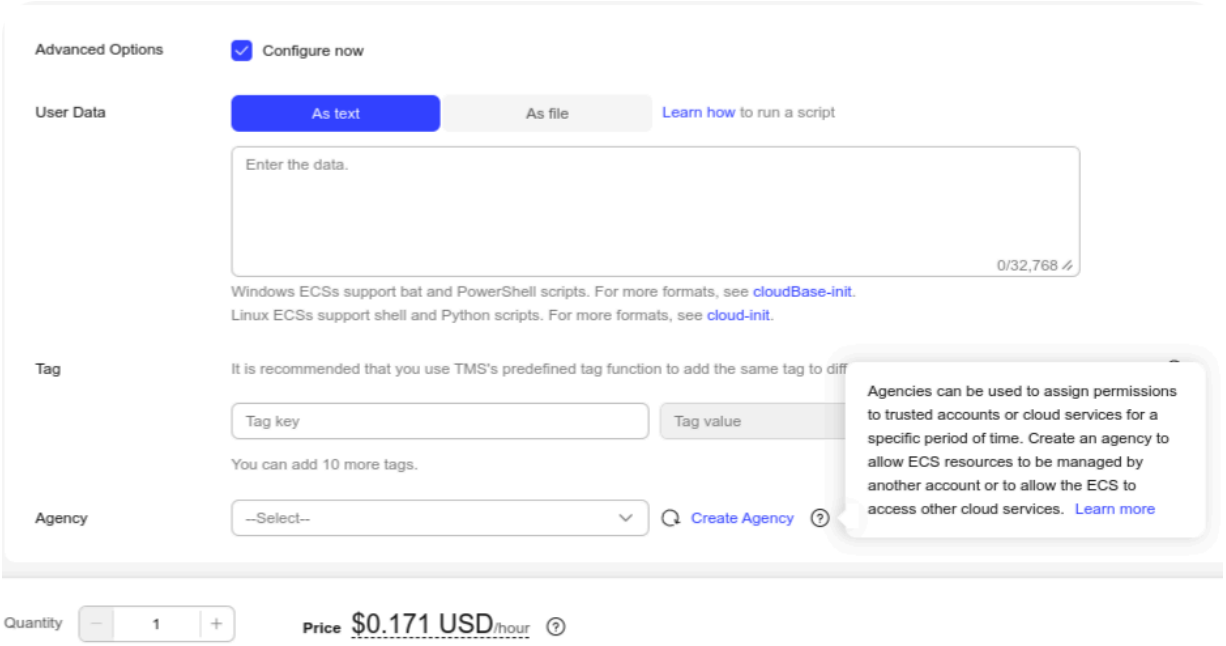


Figura 5.18: Habilitando acessos temporários

Por fim, é exibida uma tela de confirmação, com um sumário que reúne todas as configurações estabelecidas nas três etapas anteriores. Para efetivar a criação da instância, basta que o usuário concorde com as políticas estabelecidas pelo provedor, confirme a quantidade de instâncias que deseja criar e clique no botão de confirmação, disponível no canto inferior direito da tela.

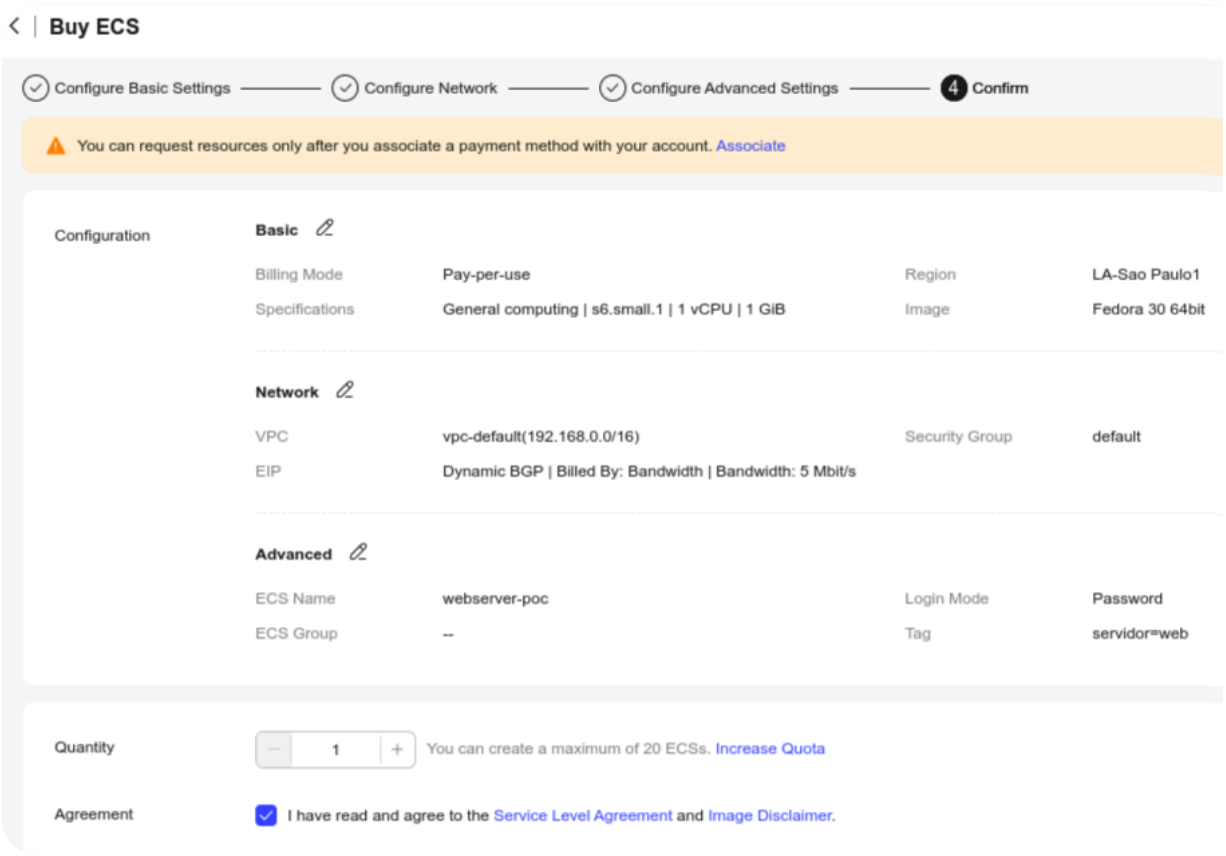


Figura 5.19: Confirmação dos parâmetros da instância

5.2 Serviços de computação em nuvem baseados em contêineres

Conforme mencionado anteriormente, dentre os serviços de computação disponibilizados pelos provedores de nuvem pública, os baseados em virtualização são os mais populares, devido à sua simplicidade e facilidade de implementação. No entanto, é importante que o usuário da tecnologia tenha em mente que processos de provisionamento ou migração de infraestrutura para ambientes de nuvem representam uma mudança de paradigma realmente significativa.

Para que todos os impactos dessa mudança possam ser sentidos de forma positiva, é necessário que as aplicações migradas ou desenvolvidas para o novo ambiente levem em conta suas particularidades e, especialmente, tirem proveito de todas elas. Aplicações com essas características recebem o nome de "*cloud native*".

Em um eventual cenário de migração de aplicações legadas para ambiente de nuvem, o processo de implantação e efetivo funcionamento tende a acontecer de forma rápida, uma vez que o cliente tenha optado pelo modelo *lift and shift*, no qual a arquitetura original é preservada, com pouca ou nenhuma mudança. A grande questão concentra-se no fato de que os módulos funcionais de um sistema monolítico são interdependentes, o que os torna fortemente acoplados, e dificulta a utilização de estratégias tipicamente adotadas em ambiente de nuvem, como a escalabilidade dinâmica, a tolerância a falhas e consequentemente, a disponibilidade.

Nesse sentido, a adoção de um modelo de desenvolvimento baseado em microsserviços é altamente recomendada para novas aplicações, sobretudo as que serão implementadas em ambiente de nuvem. Resumidamente, a tecnologia permite que cada componente de *software* pertencente à aplicação trabalhe de forma independente e especializada, daí a denominação de "microsserviço", comunicando-se com os demais via *APIs*.

Isolar cada funcionalidade da aplicação como um microsserviço especializado facilita a gerência e a manutenção ao longo do tempo. Como cada microsserviço atua de forma isolada dos demais, seu desenvolvimento, implantação e manutenção tornam-se independentes, cada um sendo executado em um contêiner específico. Essa abordagem permite que modificações de código ou de escalabilidade de um microsserviço específico sejam feitas a qualquer tempo e de forma isolada, sem afetar os demais contêineres em execução.

Além disso, em função das características de *design* de um contêiner, o eventual processo de reinício de um microsserviço ocorre na velocidade de execução de um processo, ou seja, praticamente instantâneo. Isso torna os processos de escalabilidade dinâmica, balanceamento de carga, tolerância a falhas e, consequentemente, a disponibilidade muito mais elegantes e factíveis e bem menos perceptíveis por parte do usuário final, na eventual ocorrência de um incidente.

A utilização de instâncias de virtualização, por outro lado, apesar de permitir o uso de recursos de *Auto Scaling*, apresenta um tempo de *bootstrap* significativamente maior quando comparada a um contêiner, uma vez que precisa inicializar todos os componentes do sistema operacional, além dos demais *softwares* e serviços pertinentes à aplicação que pretende disponibilizar.

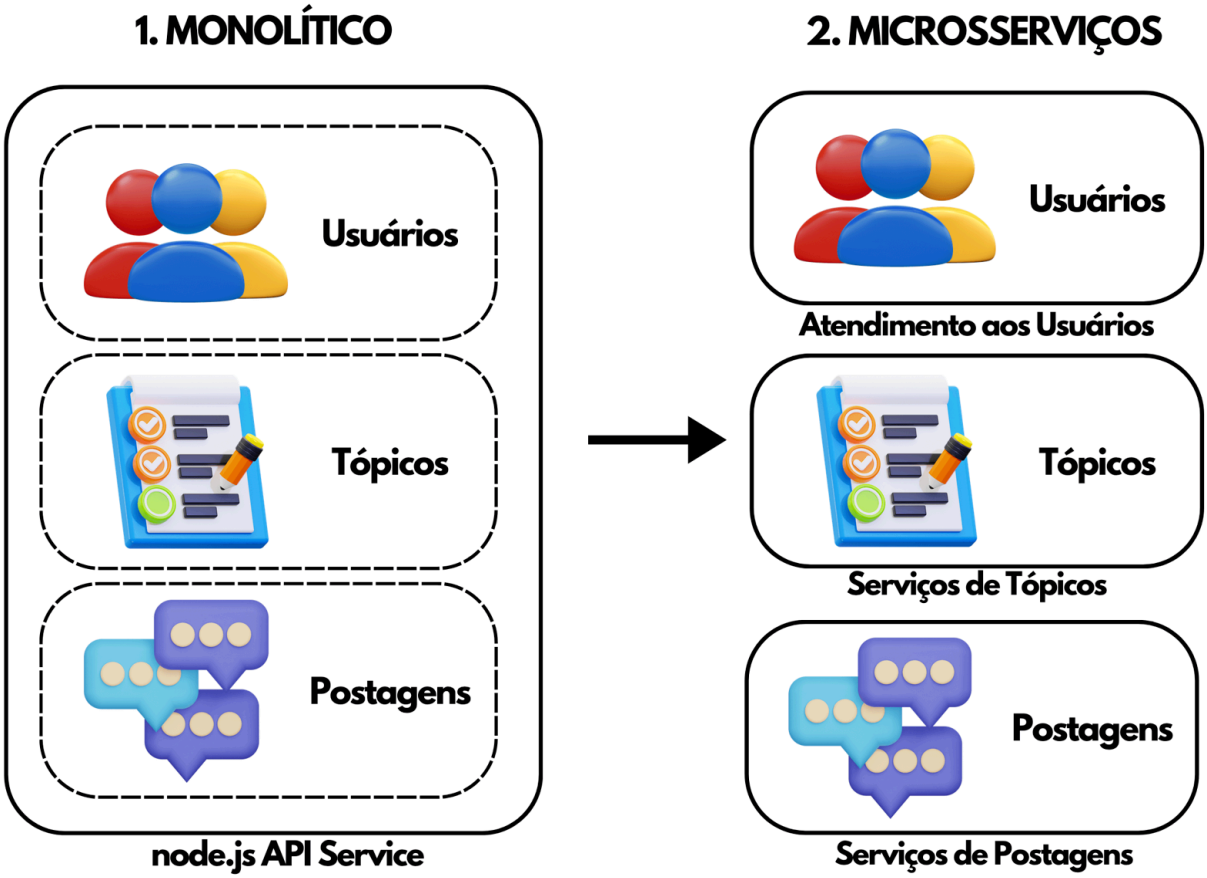


Figura 5.20: Aplicações monolíticas versus microsserviços

Obviamente, dada a relevância e complexidade do tema, não é objetivo desta obra aprofundar a discussão sobre contêineres e microsserviços. A intenção dessa seção é reconhecer a importância dessas tecnologias, sobretudo em ambientes distribuídos. A adoção de tecnologias adequadas em ambientes de nuvem — inclusive em cenários *multicloud* — pode representar, em médio ou longo prazo, o sucesso e crescimento do negócio após a adoção do novo paradigma. Por outro lado, escolhas equivocadas podem levar a um cenário de arrependimento e até à repatriação de dados.

Trabalhando com contêineres em ambiente de nuvem

Tirar proveito dos benefícios proporcionados pela utilização da arquitetura de microsserviços e, consequentemente, dos contêineres

não depende, necessariamente, da adoção do paradigma de computação em nuvem. Em ambientes de infraestrutura própria (*on-premises*) ou em estratégias de *hosting* baseadas em *datacenters* de terceiros, é perfeitamente possível implementar um ambiente de orquestração de contêineres e usufruir dos benefícios da tecnologia. No entanto, o ambiente de nuvem apresenta diferenciais que merecem atenção.

Suponhamos que um cliente tenha optado por migrar uma aplicação hipotética, baseada em microsserviços, para o ambiente de nuvem. Na etapa inicial do processo, é preciso que o usuário escolha como deseja realizar o provisionamento da infraestrutura necessária. Em um primeiro momento, a título de prova de conceito, é natural que o cliente opte por utilizar uma instância de virtualização baseada em uma imagem contendo uma distribuição *GNU/Linux*, justamente para a instalação da "*container engine* desejada. Uma vez provisionada a infraestrutura, ficam evidentes os benefícios obtidos com a utilização da tecnologia de contêineres, tais como:

- constatação da efetiva reprodutibilidade do ambiente necessário para execução da aplicação, de forma imediata: tal característica é tida como chave na utilização da tecnologia de contêineres, pois elimina situações outrora comuns, como a máxima do "na minha máquina funciona", típica da era que antecedeu a adoção da cultura *DevOps* em ambiente de produção;
- possibilidade de utilização de praticamente todo o portfólio de ferramentas disponibilizado pelo provedor de nuvem em favor da utilização dos microsserviços: independentemente da arquitetura escolhida pelo cliente, os serviços interagem entre si como um grande "lego", permitindo que diferentes funcionalidades — como as citadas na seção de serviços de computação baseados em virtualização — sejam incorporadas com facilidade em uma eventual transição tecnológica. Um exemplo é o processo de refatoração ou reengenharia de uma

aplicação, a princípio monolítica, que passa a utilizar uma nova abordagem, baseada em microsserviços.

Uma vez concluída a etapa de experimentação, é necessário que a aplicação seja provisionada em ambiente de produção, para que finalmente possa ser utilizada. Para isso, o uso de um sistema de orquestração de contêineres torna-se um requisito obrigatório. Dentre as soluções disponíveis voltadas para essa finalidade, a mais utilizada é, sem dúvidas, o *Kubernetes*, ou simplesmente, *k8s*. Como dito anteriormente, é perfeitamente possível utilizá-lo fora do ambiente de nuvem, no entanto, o usuário terá de lidar com a complexidade envolvida na gestão de um *cluster* de orquestração de contêineres. Temos então uma das maiores vantagens da migração para ambiente de nuvem: a disponibilidade de serviços de orquestração de contêineres totalmente gerenciados pelo provedor de nuvem pública.

A *Huawei Cloud* disponibiliza o *Cloud Container Engine (CCE)* como sua solução de *Kubernetes as a Service*, sendo um dos primeiros provedores de serviços *Kubernetes* certificados pela *Cloud Native Computing Foundation (CNCF)*. A utilização do serviço resume-se a alguns passos simples. Inicialmente, o usuário define os parâmetros necessários para a criação do *cluster*, podendo optar por um *cluster* padrão ou pela versão "turbo", que oferece especificações aprimoradas. Na sequência, realiza o *upload* de imagens, *templates* e arquivos *yaml* eventualmente necessários para o provisionamento dos componentes da aplicação.

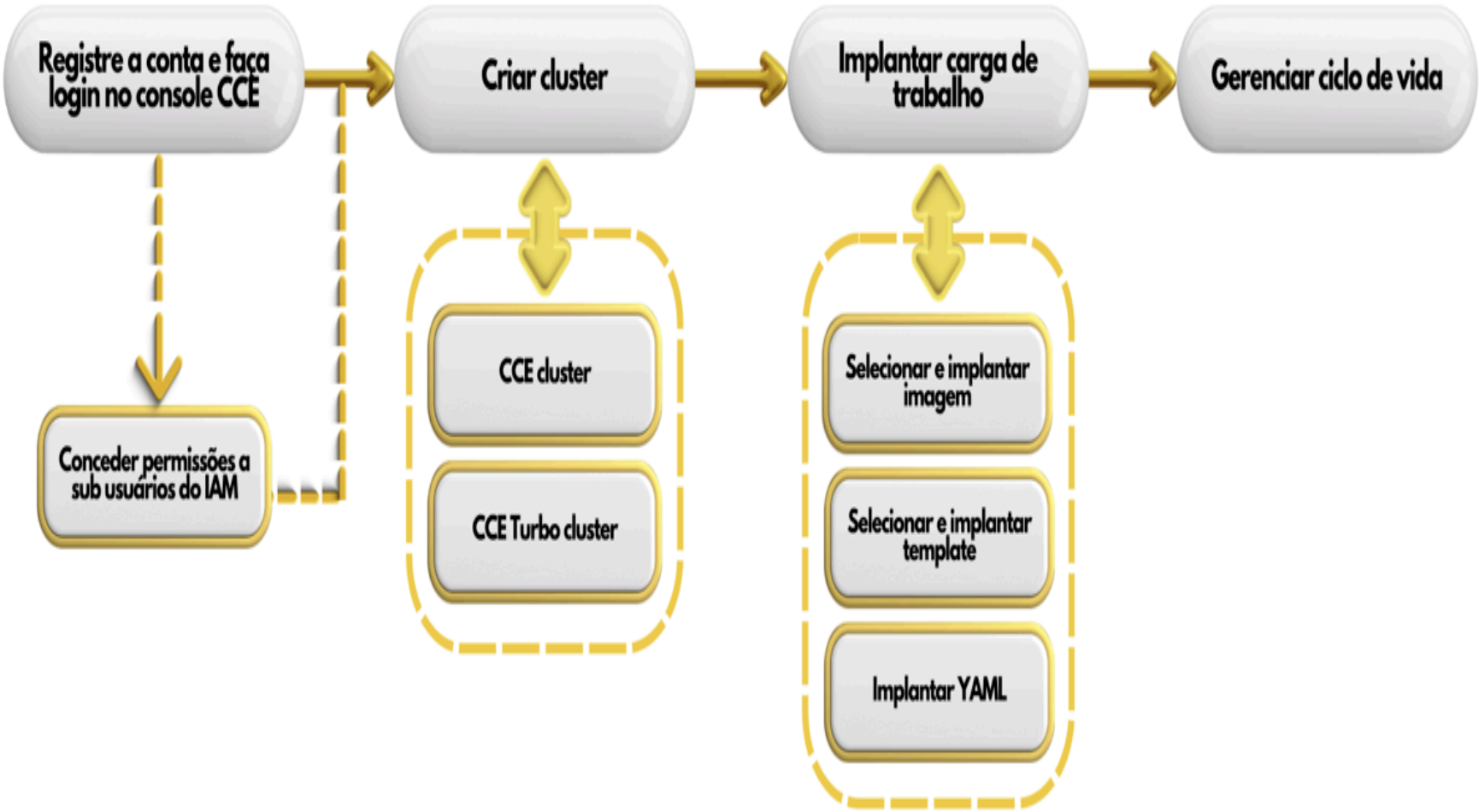

Registre a conta e faça login no console CCE
Criar cluster
Implantar carga de trabalho
Gerenciar ciclo de vida
Conceder permissões a sub usuários do IAM
CCE cluster
CCE Turbo cluster
Selecionar e implantar imagem
Selecionar e implantar template
Implantar YAML

Figura 5.21: Como utilizar o Cloud Container Engine (CCE)

A utilização de uma ferramenta de orquestração de contêineres totalmente gerenciada pelo provedor de nuvem garante que a aplicação do cliente desfrute, de forma automática e nativa, de recursos de balanceamento de carga, elasticidade e alta disponibilidade, com precificação *pay-per-use*.

Nesse sentido, aplicações construídas com a arquitetura de microsserviços podem aproveitar não somente os recursos disponibilizados por soluções de *kubernetes* como serviço, ofertada pelo CCE, mas também se integrar às demais ferramentas do provedor, incluindo balanceadores de carga, funcionalidades de segurança e outros serviços, conforme detalhado a seguir.

Principais ajustes:

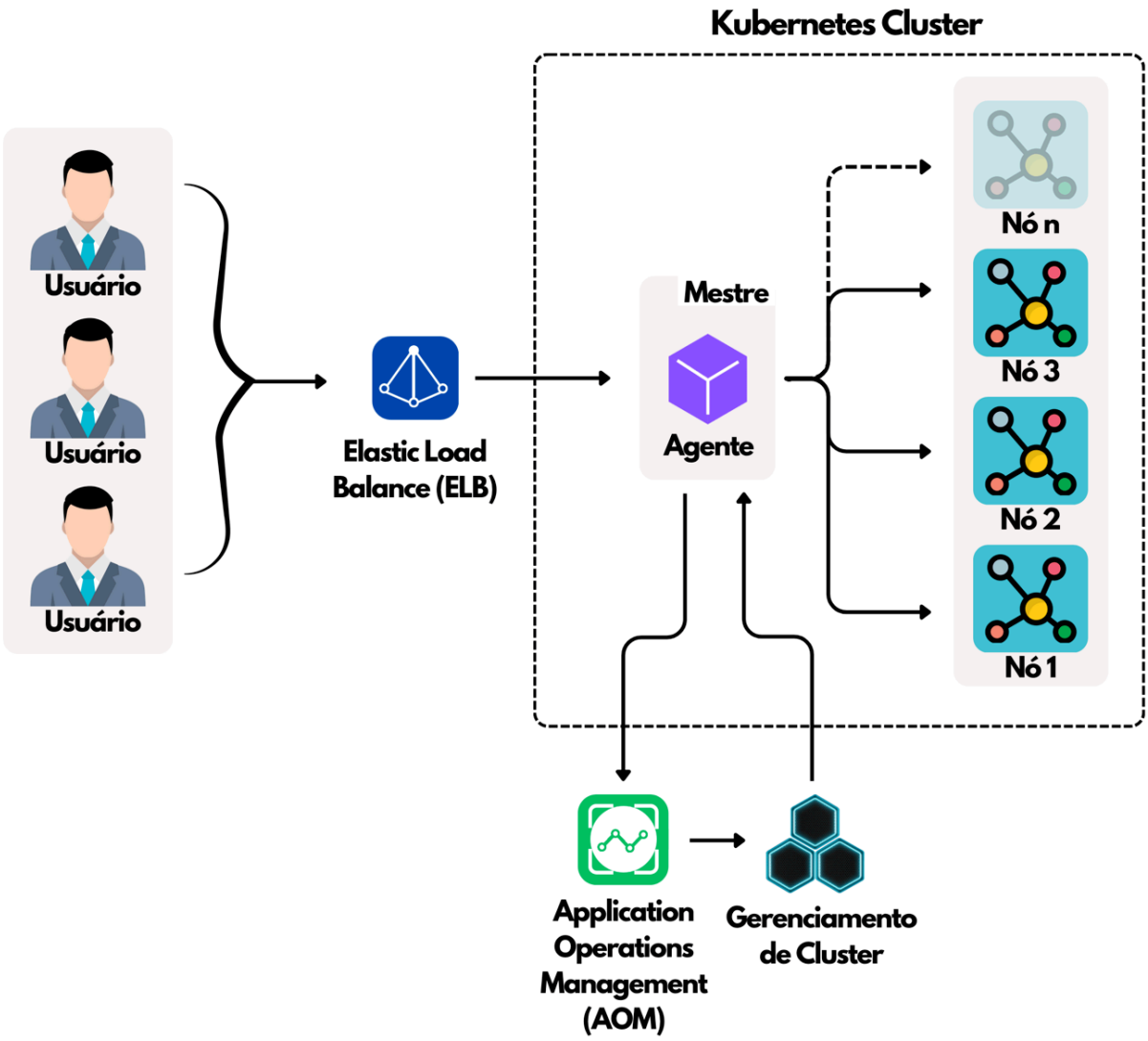


Figura 5.22: Auto Scaling

Times de desenvolvimento contam com *pipelines* de integração contínua e entrega contínua, com ambientes de teste, homologação e produção, além de um serviço privado de *registry* denominado *Software Repository for Container (SWR)*, dedicado ao armazenamento das imagens de contêineres utilizadas nas aplicações provisionadas no *cluster*.

A utilização de um repositório de imagens, também denominado como *registry*, dentro do ambiente de nuvem, possibilita baixa latência e, consequentemente, maior performance na obtenção de

imagens, maior controle frente ao permissionamento dos recursos disponibilizados, bem como a economicidade no trânsito de dados, uma vez que estes são distribuídos aos serviços de computação existentes nas *pipelines* internamente, dentro da infraestrutura do provedor de nuvem.

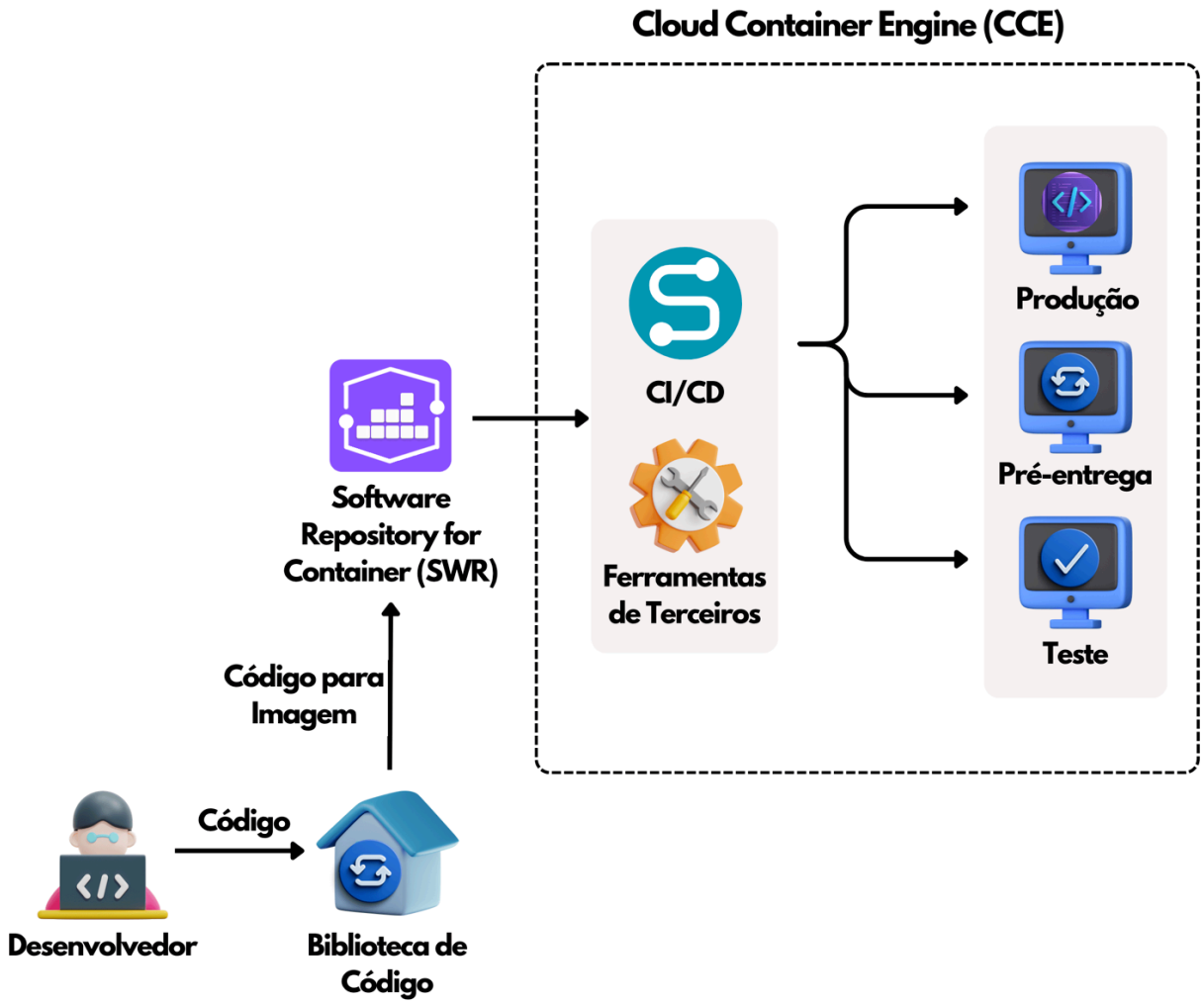


Figura 5.23: DevOps

Além disso, o *CCE* pode ser utilizado como solução para ambientes de nuvem híbrida. Nesse tipo de cenário, a infraestrutura provisionada na nuvem pode ser utilizada, por exemplo, somente sob demanda, de modo a complementar o poder computacional da infraestrutura legada *on-premises* quando necessário. De mesma

maneira, pode também servir como solução de alta disponibilidade, inclusive com aplicação *multicloud*.

A imagem a seguir demonstra a construção de uma arquitetura de nuvem híbrida, que utiliza um *cluster* de orquestração de contêineres com nós distribuídos entre o ambiente de nuvem e a infraestrutura própria mantida pelo cliente. Essa topologia permite compreender, de forma clara, a versatilidade e as possibilidades de integração entre as tecnologias já existentes nas organizações e as soluções equivalentes disponibilizadas pelo paradigma de nuvem.

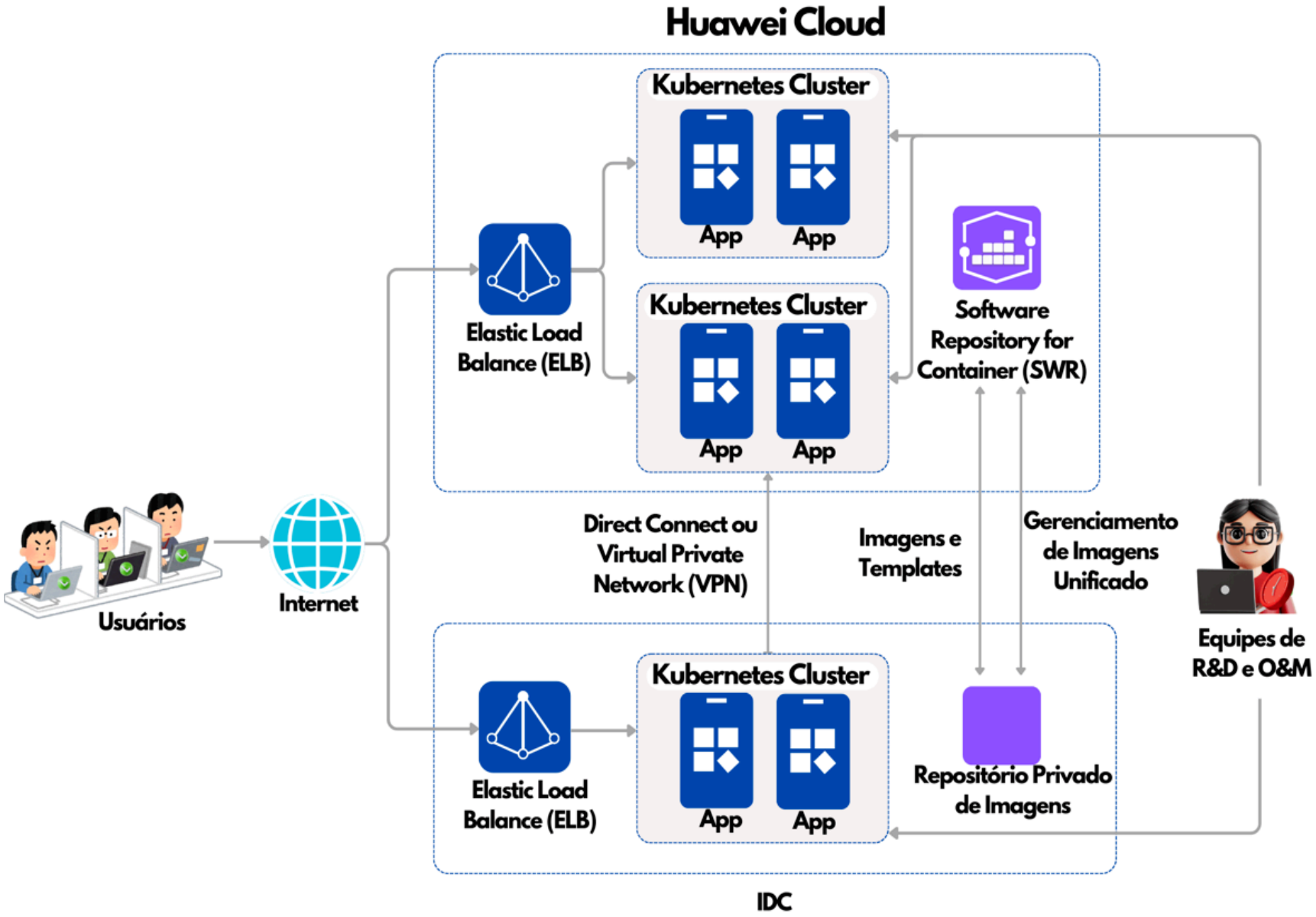

Huawei Cloud
Kubernetes Cluster
App
App
Kubernetes Cluster
App
App
Elastic Load Balance (ELB)
Software Repository for Container (SWR)
Usuários
Internet
Direct Connect ou Virtual Private Network (VPN)
Imagens e Templates
Gerenciamento de Imagens Unificado
Equipes de R&D e O&M
Kubernetes Cluster
App
App
Elastic Load Balance (ELB)
Repositório Privado de Imagens
IDC

Figura 5.24: Nuvem híbrida

A computação em nuvem evoluiu de uma realidade inicial, fortemente apoiada em tecnologias de virtualização, passando por um estágio intermediário focado nos recursos tipicamente definidos por *software*, até alcançar um cenário atual em que o ponto alto da tecnologia é centrado nas aplicações, especialmente as denominadas "*cloud native*". Nesse contexto, é fundamental destacar a importância dos serviços de computação baseados em contêineres para assegurar a continuidade de negócios em uma organização. Tenha isso sempre em mente.

5.3 Serviços de computação em nuvem baseados em eventos

A crescente demanda pela utilização de serviços de diferentes categorias disponibilizados pelos provedores de computação em nuvem proporcionou o surgimento de novas tecnologias promissoras, dentre as quais destaca-se a computação sem servidor, ou *serverless*. A relevância do modelo proposto é notória, tendo em vista que, rapidamente, passou a ser referenciada na literatura como uma nova e inovadora modalidade de serviço, denominada *Function as a Service*, ou simplesmente *FaaS*. A proposta desse serviço é fornecer ao cliente uma interface de gerenciamento que permita o provisionamento de códigos, executados mediante determinada circunstância ou evento, em uma infraestrutura sem servidores.

Um exemplo clássico e muito utilizado para exemplificar um caso de uso factível para utilização da tecnologia consiste na monitoração de um *bucket* de origem, fornecido pelo serviço de armazenamento de objetos do provedor de nuvem, que atua como repositório de imagens enviadas pelo cliente. Vamos imaginar que o cliente faça o *upload* dessas imagens por meio de uma aplicação móvel instalada em seu aparelho celular, capturadas diretamente pela câmera do

dispositivo na resolução máxima disponível. Supondo que o usuário deseje criar *thumbnails* para otimizar a visualização e a busca dos arquivos enviados, seria de grande valia automatizar todo o processo.

Para tanto, utilizaremos o *FunctionGraph*, serviço de computação *serverless* disponibilizado pela *Huawei Cloud*, que permite ao usuário vincular a ocorrência de um evento — nesse caso, a gravação de dados em um *bucket* do *OBS* — como gatilho para a execução de um código disponibilizado pelo cliente. Esse código gerará a miniatura do arquivo cujo *upload* acabou de ser realizado em um segundo *bucket* de destino.

O detalhe é que todo esse processo é realizado sem que o usuário precise provisionar ou gerenciar servidores, pois os serviços de computação sem servidor são totalmente gerenciados pelo provedor de nuvem, o que garante a alta disponibilidade e a escalabilidade dinâmica da infraestrutura necessária de acordo com o aumento ou diminuição da demanda real. Cabe ao usuário disponibilizar o código-fonte da função a ser utilizada e definir as condições desejadas para sua execução. O modelo de cobrança adotado no serviço utiliza como métrica para realizar a precificação o número de execuções da função em questão. Dessa maneira, o cliente não é cobrado quando o código não está em funcionamento.

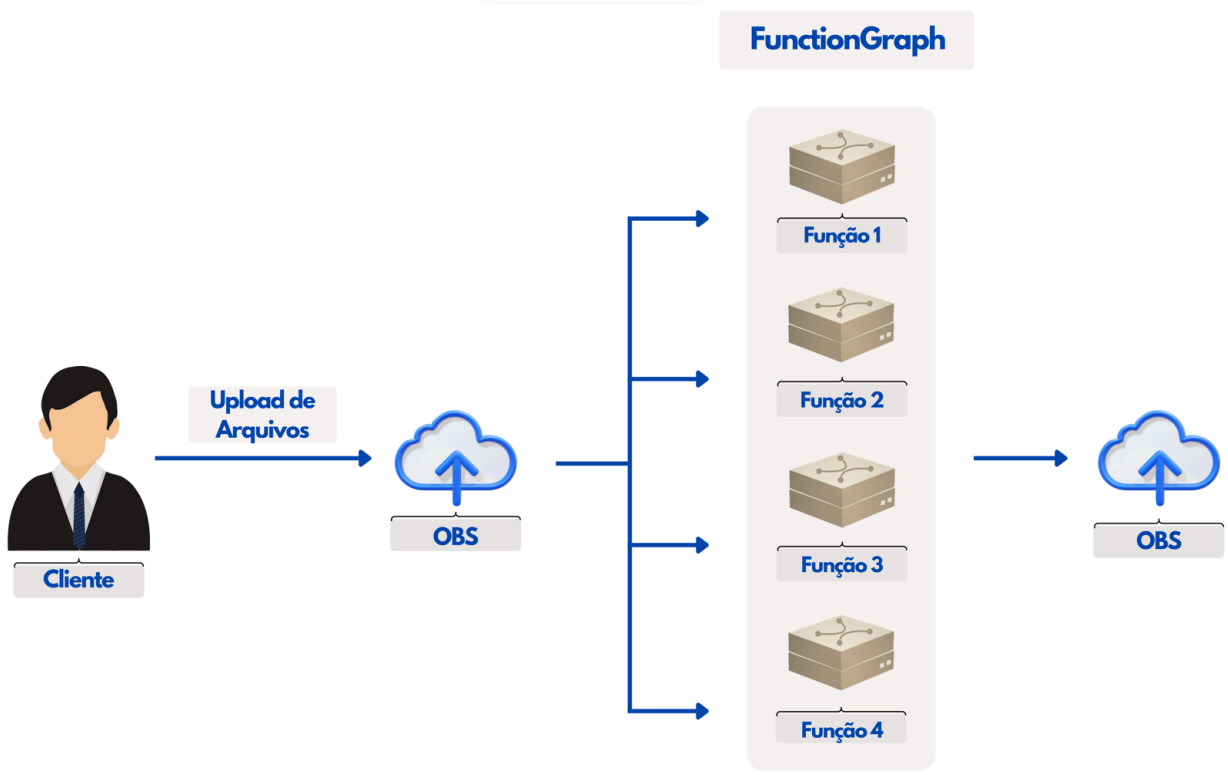


Figura 5.25: Processamento de arquivos em tempo real

Evidentemente, esse modelo de serviço não é adequado para qualquer tipo de aplicação. Apesar de seu uso limitado a finalidades específicas, os provedores de nuvem têm ampliado o número de serviços que disponibilizam eventos de ativação compatíveis com a tecnologia, bem como o suporte a diferentes linguagens de programação, o que aumenta significativamente os possíveis cenários de aplicação desta modalidade de serviço de computação, cada dia mais popular.

5.4 Considerações importantes

Vale a pena reiterar que alternativas de serviços de computação, com funcionalidades equivalentes às demonstradas a título de referência, estão disponíveis em todos os principais provedores de

nuvem pública. Os parâmetros utilizados, bem como a lógica de funcionamento de cada um, seguem os mesmos princípios apresentados no presente capítulo.

Por fim, é importante ter em mente que as instâncias reservadas trazem benefícios que vão além da atraente precificação dos serviços de computação. Como o próprio nome sugere, as instâncias reservadas garantem que o poder computacional adquirido pelo cliente esteja disponível para uso a qualquer tempo, especialmente em datas comerciais com alto volume de solicitações de operações de *scale out* na modalidade sob demanda, como a *black friday*, por exemplo.

Provedor	***Containers***	**Computação *Serverless***	**Virtualização**
Alibaba Cloud	ECI	Function Compute	ECS
Amazon Web Services	ECS, Fargate	Lambda	EC2
Google Cloud Platform	Cloud Build, Cloud Run	App Engine	Compute Engine
Huawei Cloud	CCE	Function Graph	ECS
IBM Cloud	IBM Cloud Kubernetes Service	IBM Cloud Code Engine	IBM Cloud Virtual Servers
Microsoft Azure	Instâncias de Contêiner do Azure	Azure Functions	Máquinas Virtuais do Azure
Oracle Cloud Infrastructure	Container Instances	Cloud Functions	OCI Compute

Questionário online

Figura 5.26: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 5.27: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo

Capítulo 6
Serviços de armazenamento

A utilização de serviços de armazenamento de dados, por motivos óbvios, é uma demanda inevitável. Seu planejamento, seja em processos de implantação ou de migração de infraestrutura para ambientes em nuvem, torna-se essencial e estratégico, pois é necessário refletir sobre seus impactos e implicações no uso cotidiano de curto, médio e longo prazo.

Inicialmente, é necessário identificar quais dados, de fato, serão migrados para o provedor de nuvem, bem como quais devem permanecer *on-premises*, sob custódia da organização. Esse tipo de estudo é fundamental, pois, em um primeiro momento, é natural que o cliente considere a possibilidade de migrar integralmente os seus dados. No entanto, essa decisão não é tão simples quanto parece.

Um ponto de partida importante são as características relacionadas ao uso dos dados em questão. Vale lembrar que, durante o ciclo de vida de um tipo hipotético de dado, uma mesma informação pode apresentar diferentes níveis de prioridade de acesso, a saber:

- Em um primeiro momento, as informações podem ser consumidas internamente, por meio de processos de consulta realizados inúmeras vezes por diferentes colaboradores de uma organização — ou, ainda, por clientes externos, a depender das características dos dados e das regras de negócios aplicadas —, gerando um alto fluxo de trânsito de informações;

- Passado algum tempo, esse mesmo conjunto de dados tende a perder relevância, muito embora continue sendo necessário. Nesse contexto, a informação permanece disponível para consulta; no entanto, os dados são armazenados e mantidos em

repouso, com baixo trânsito de informações, mas ainda ocupando espaço no volume de armazenamento;
- Na etapa seguinte, a permanência das informações no ambiente de produção torna-se desnecessária. Para fins de conformidade, é mandatório que os dados em questão sejam arquivados. A partir desse momento, não estamos mais lidando com dados prioritários, sendo eles acessados somente sob demanda, de forma assíncrona;

- Por fim, após um período de tempo preestabelecido pela política de segurança da informação da organização, os dados podem, finalmente, ser excluídos de forma definitiva.

Dado o cenário supracitado, podemos observar dois indicadores que desempenham papéis fundamentais para a elaboração de uma estratégia assertiva de armazenamento de dados em ambiente de nuvem: a quantidade de dados que devem ser efetivamente armazenados e o volume de transferência de dados de saída necessário para que possam ser consumidos.

A primeira métrica é de fácil entendimento, uma vez que o principal fator de precificação dos serviços de armazenamento leva em conta a quantidade de *gigabytes (GB)* utilizados pelo cliente. A grande questão está no segundo indicador, uma vez que os provedores não cobram pelo trânsito de dados de entrada necessário para que os arquivos sejam armazenados na nuvem. Isso pode levar o cliente, em um primeiro momento, a considerar apenas os custos de armazenamento e a migrar todos os seus dados para o novo ambiente. No entanto, o tráfego de saída necessário para, efetivamente, servir aos clientes — e, eventualmente, à própria organização — é tarifado. Esse fator certamente impactará as despesas mensais destinadas ao provisionamento e à manutenção da infraestrutura e nem sempre é considerado como deveria.

Os serviços de armazenamento disponibilizados pelos provedores de nuvem, com foco em três funcionalidades essenciais, serão detalhados nas seções a seguir:

- Serviços de armazenamento em nuvem baseados em blocos;
- Serviços de armazenamento em nuvem baseados em arquivos;
- Serviços de armazenamento em nuvem baseados em objetos.

6.1 Serviços de armazenamento em blocos

Os serviços de armazenamento em blocos são utilizados para a criação de volumes lógicos, cuja principal finalidade é a persistência de dados. Para que isso seja possível, os volumes criados pelos clientes com os serviços são utilizados em conjunto com serviços de computação de diferentes tipos, como instâncias de virtualização, servidores dedicados e afins, podendo atuar como uma unidade de inicialização ou como unidade adicional de armazenamento.

O tipo e o desempenho do dispositivo de armazenamento utilizado também podem ser determinados pelo cliente. Estão disponíveis para uso discos mecânicos tradicionais (HDs), que, em geral, devem ser destinados aos volumes secundários, bem como discos de estado sólido (SSDs), mais performáticos e utilizados para os discos de inicialização, que contêm o sistema operacional da imagem escolhida na configuração do serviço de computação.

Ainda dentro das categorias de disco, é possível escolher entre diferentes especificações de desempenho, referentes, por exemplo, à quantidade de operações de leitura e escrita por segundo (IOPS) que o volume é capaz de suportar. A escolha de uma unidade de armazenamento com maior capacidade de operações por unidade de tempo possibilita um uso mais eficiente em aplicações que demandem alta vazão de dados, múltiplos acessos simultâneos ou gravação recorrente de *streaming* de dados, entre outros.

Outro fator relevante é que os volumes criados contam com recurso de escalabilidade elástica. Dessa forma, as unidades podem ser redimensionadas a qualquer tempo, aumentando a capacidade de

armazenamento sempre que necessário. Vale lembrar que, ao aumentar o tamanho do volume no painel de gerenciamento do serviço de armazenamento em blocos utilizado, é necessário também promover o redimensionamento da partição lógica no disco no sistema operacional da instância para que o espaço adicional seja efetivamente reconhecido e utilizado. Não é possível reduzir o tamanho de uma unidade de armazenamento, somente aumentá-lo.

Para fins de segurança, é possível encriptar os dados do volume de armazenamento e assim adicionar uma camada de segurança aos dados armazenados em repouso. Os provedores de nuvem também oferecem serviços de gerenciamento de chaves criptográficas, liberando o cliente dessa responsabilidade. Por padrão, os discos não são encriptados, cabendo ao usuário tomar essa decisão.

Um dos recursos mais interessantes dos serviços de armazenamento em blocos é a possibilidade de criação de *backups* e *snapshots*. Esses conceitos são frequentemente confundidos e, por vezes, tratados como um recurso único, embora possuam diferenças significativas, a saber:

- O processo de *backup* tem por objetivo, a rigor, manter cópias de arquivos importantes em repositórios confiáveis de dados para que possam ser recuperados, se necessário. Podem ser realizados *backups* totais ou incrementais: o primeiro copia todos os arquivos selecionados para o destino de uma única vez; o segundo, após a realização do primeiro, executa, dentro de uma determinada periodicidade, somente a porção modificada dos arquivos alterados desde o último *backup* total ou parcial;

- Os *snapshots* representam imagens completas de um volume de armazenamento que, em termos práticos, funcionam como uma "fotografia" do sistema de arquivos em um dado momento, ou *timestamp*. Em geral, costumam ser criados em momentos estratégicos que antecedem operações importantes, possibilitando ao administrador de sistemas restaurar

rapidamente uma estação de trabalho a uma condição funcional anterior, após a instalação de um *software* ou uma atualização de sistema operacional malsucedida, por exemplo.

Além disso, é possível monitorar as unidades de armazenamento por meio de diferentes métricas, com o objetivo de mitigar problemas relacionados à manutenção desses tipos de recursos, especialmente aqueles ligados à performance, à falta de espaço em disco, entre outros fatores.

Dentre as principais características de um serviço de armazenamento em blocos, destaca-se o fato de que a utilização de um volume de armazenamento pode ser feita por apenas um dispositivo final por vez, não sendo possível o consumo simultâneo dos dados armazenados na unidade por duas ou mais instâncias de virtualização, por exemplo.

A *Huawei Cloud* utiliza o *Elastic Volume Service (EVS)* para o provisionamento desse tipo de serviço. O acesso aos recursos de armazenamento em blocos pode ser feito diretamente pela interface do serviço, especialmente para fins de gerenciamento e manutenção. Também é possível criar os recursos sob demanda, por meio dos assistentes disponíveis nos serviços de provisionamento de instâncias de virtualização ou servidores dedicados, conforme vimos no capítulo anterior.

6.2 Serviços de armazenamento em arquivos

Ao contrário dos serviços de armazenamento de dados em blocos, os baseados em arquivos foram criados para permitir o uso simultâneo por várias instâncias, o que constitui sua principal característica. Essa modalidade de armazenamento utiliza, em sua infraestrutura, sistemas de arquivos distribuídos clássicos, como o *Network File System (NFS)* e o *Server Message Block (SMB)*, a

depender do provedor de nuvem utilizado, permitindo o compartilhamento de recursos via rede. Para consumir os dados, é possível utilizar os principais serviços de computação disponibilizados pelo provedor, sejam eles baseados em virtualização, contêineres ou servidores dedicados.

No portfólio de produtos da *Huawei Cloud*, o recurso responsável pela implementação do serviço de armazenamento em arquivos denomina-se *Scalable File Service*, ou simplesmente *SFS*. Para iniciarmos os passos necessários à implementação, é importante entender que o processo, como um todo, é composto por pelo menos duas etapas, independentemente do provedor de nuvem utilizado:

- Criação do sistema de arquivos a ser compartilhado via rede;
- Montagem do sistema de arquivos compartilhado em um dos clientes que vão utilizá-lo.

Para iniciar o processo de criação do sistema de arquivos a ser compartilhado, basta clicar no botão *"Create File System"*, localizado no canto superior direito do console de gerenciamento *web* pertinente ao *SFS*.

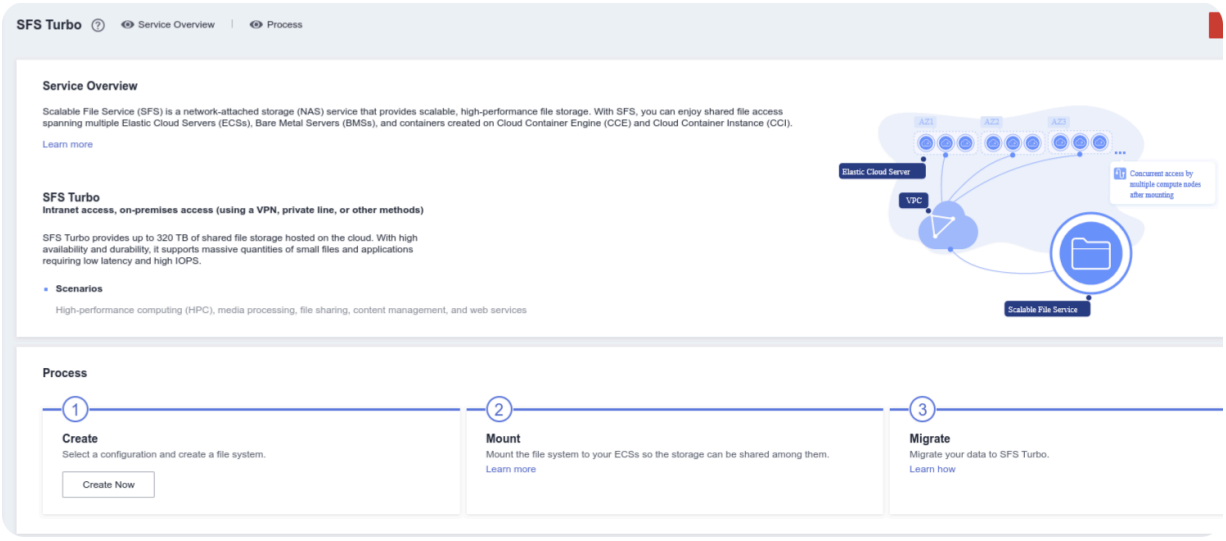


Figura 6.1: Tela inicial da interface do Scalable File Service - SFS

Em seguida, o usuário deve preencher as informações de provisionamento da infraestrutura de acordo com as necessidades do projeto. Inicialmente, o assistente de configuração solicita a região e a zona de disponibilidade em que o recurso deverá ser criado, bem como o nome do projeto ao qual ele deverá estar associado.

O usuário deve escolher também se o projeto em questão é voltado para computação de alto desempenho *(HPC)* ou para um projeto de propósito geral *(General*). A escolha entre uma ou outra opção impactará diretamente a disponibilidade de opções referentes à performance de transferência de dados, à capacidade de armazenamento e, consequentemente, na precificação. A título de ilustração, utilizaremos a opção *File System Type: General*, com capacidade inicial de armazenamento de 500 *GB*, pertencente à classe de armazenamento *"Performance"* e utilizando o protocolo *NFS*.

As demais opções determinam em qual *VPC* e em qual sub-rede o recurso será alocado; se o serviço de *Cloud Backup and Recovery*, voltado para recuperação de desastres, deve ser utilizado; além de opções avançadas e opcionais relacionadas à criptografia, grupos de segurança e *tags*, bem como o nome que deverá ser associado ao recurso que está prestes a ser criado. Para proceder com a criação do sistema de arquivos compartilhado, basta clicar no botão *"Create Now*"*, localizado no canto inferior direito da interface.

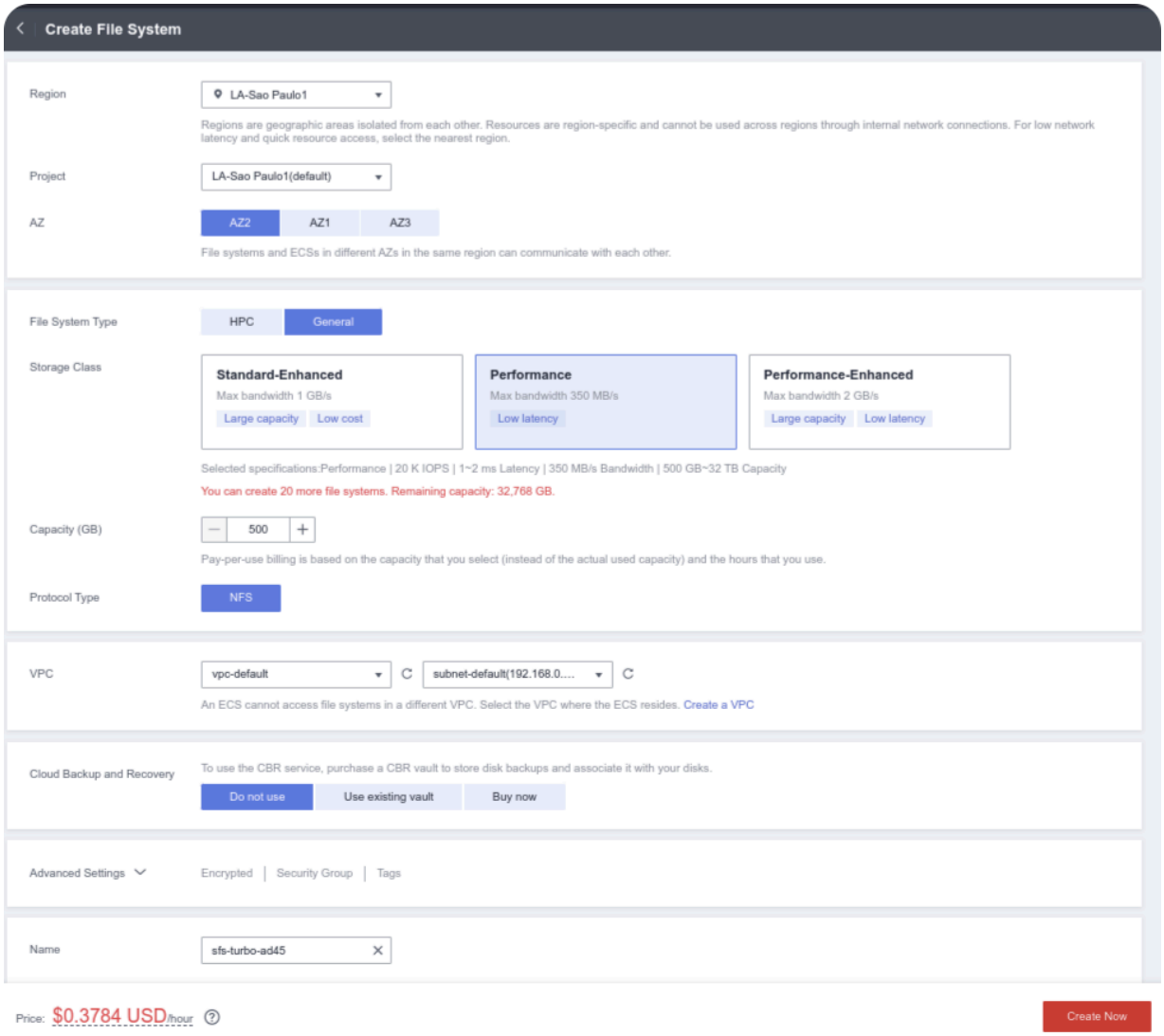


Figura 6.2: Tela inicial da interface do Scalable File Service - SFS

Uma vez criado o sistema de arquivos a ser compartilhado, é necessário criar os pontos de montagem em cada uma das instâncias onde se deseja acessar o recurso. Os serviços de computação utilizados para consumir o sistema de arquivos compartilhado, conforme mencionado anteriormente, podem ser baseados em diferentes tecnologias. A configuração de cada estação de trabalho resume-se à instalação de um cliente *NFS*, bem como à definição dos parâmetros referentes ao compartilhamento criado no passo inicial, que, em geral, são disponibilizados pelo provedor ao final do processo. As instâncias nas quais os pontos de montagem

foram criados podem estar provisionadas em diferentes zonas de disponibilidade ou, até mesmo, na infraestrutura *on-premises* do cliente.

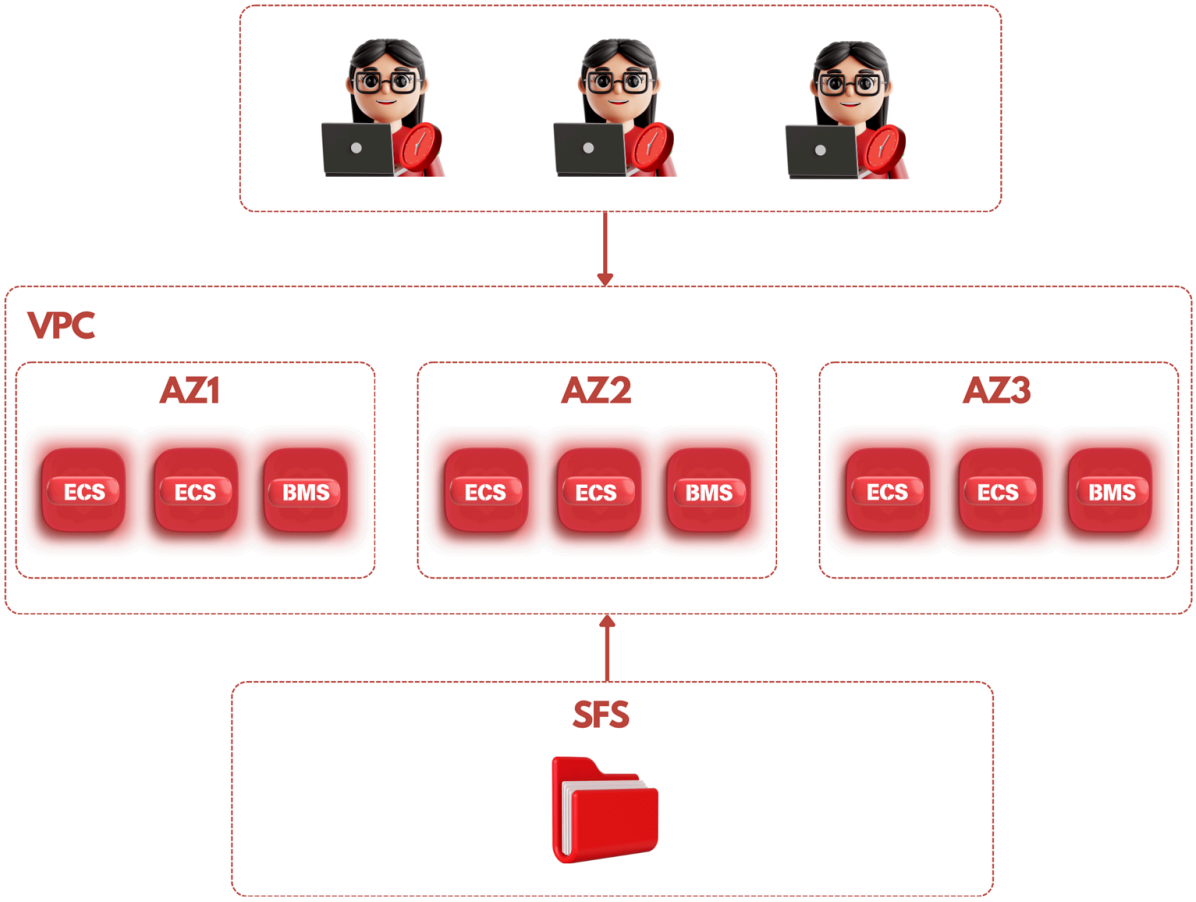


Figura 6.3: Scalable File Service - SFS, da Huawei Cloud, utilizado como ferramenta de conteúdo para compartilhamento de arquivos entre diferentes AZs

Outro uso comum dos serviços de armazenamento baseados em compartilhamento de arquivos é servir como repositório centralizado de conteúdos estáticos para serviços e aplicações que utilizam escalabilidade horizontal e balanceamento de carga, como servidores *web*, por exemplo.

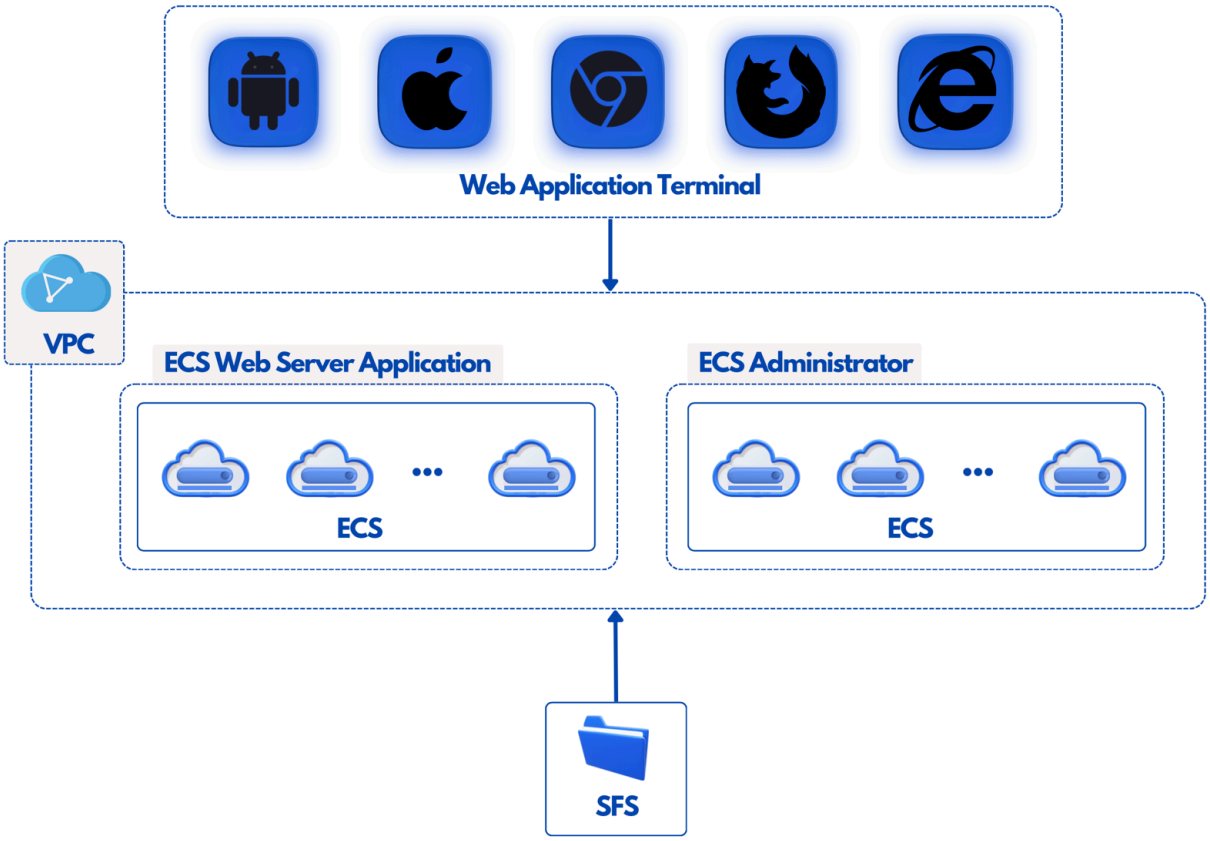


Figura 6.4: Scalable File Service - SFS, da Huawei Cloud, utilizado como repositório centralizado de conteúdos estáticos para aplicações web

Caso o espaço alocado ao compartilhamento se torne insuficiente com o passar do tempo, é possível redimensioná-lo a qualquer momento, permitindo que as alterações sejam refletidas imediatamente nos clientes, sem a necessidade de interrupção do serviço.

6.3 Serviços de armazenamento de objetos

Os serviços que tivemos a oportunidade de conhecer nas seções anteriores podem ser facilmente relacionados a necessidades cotidianas de armazenamento de dados que qualquer usuário enfrenta em sua organização, como a gravação de dados em disco

— seja no próprio dispositivo ou em um diretório remoto compartilhado na rede. Guardadas as devidas proporções, tratam-se de casos de uso comuns em um ambiente de rede local, que continuam a ser necessários, independentemente de onde a infraestrutura esteja provisionada.

Evidentemente, a utilização desses cenários em ambiente de nuvem traz benefícios intrínsecos à tecnologia, em especial os relacionados à precificação, agilidade de implementação, escalabilidade dinâmica, autenticação, autorização, bem como aos demais recursos relativos à segurança, como as políticas de acesso, grupos de segurança, *ACLs*, entre tantos outros recursos disponíveis. Ao provisionar infraestrutura em ambiente de nuvem, naturalmente surgem novas oportunidades, juntamente a novas soluções, mais robustas e versáteis. Em tempos em que tecnologias como *big data*, *IoT* e *Machine learning* tornam-se cada dia mais necessárias em todo tipo e tamanho de negócio, é esperado que os serviços básicos que dão suporte a tais tecnologias acompanhem o dinamismo dessa tendência.

O armazenamento de objetos traz inovações significativas à categoria de serviço que representa, pois se trata de um tipo de serviço totalmente gerenciado pelo provedor de serviço, com capacidade de armazenamento virtualmente ilimitado, capaz de atender a altas demandas de acesso simultâneo, com alta disponibilidade, diferentes modelos de precificação e baixa latência.

Sendo assim, dentre as opções disponíveis na categoria a que pertence, os serviços voltados ao armazenamento de objetos são a escolha natural quando o usuário deseja arquivar grandes volumes de dados. Na *Huawei Cloud*, o serviço responsável por esse tipo de funcionalidade é o *Object Storage Service (OBS)*. A arquitetura do serviço está estruturada em dois componentes principais, os *buckets* e os objetos:

- O primeiro atua como um repositório de armazenamento de objetos associado a determinado projeto ou demanda

específica. Por definição, recebe nomes únicos e, por esse motivo, recomenda-se a utilização de domínios *DNS* na composição de sua nomenclatura. Por motivos de segurança, tem visibilidade definida como privada por padrão; no entanto, é possível tornar esse atributo público, caso necessário;

- Os objetos, por sua vez, são entidades de armazenamento que populam os *buckets*. Diferem-se dos arquivos tradicionais por possuírem três atributos principais, denominados *"key", "metadata" e "data"*. O primeiro atributo é responsável pela definição do nome do objeto. O segundo é composto por um conjunto de pares do tipo chave-valor que auxiliam a identificação do objeto, sendo que alguns são atribuídos automaticamente nos processos de criação e manipulação dos objetos, enquanto outros são definidos pelo usuário, de forma personalizada. O terceiro e último atributo, como o próprio nome sugere, refere-se aos dados propriamente ditos.

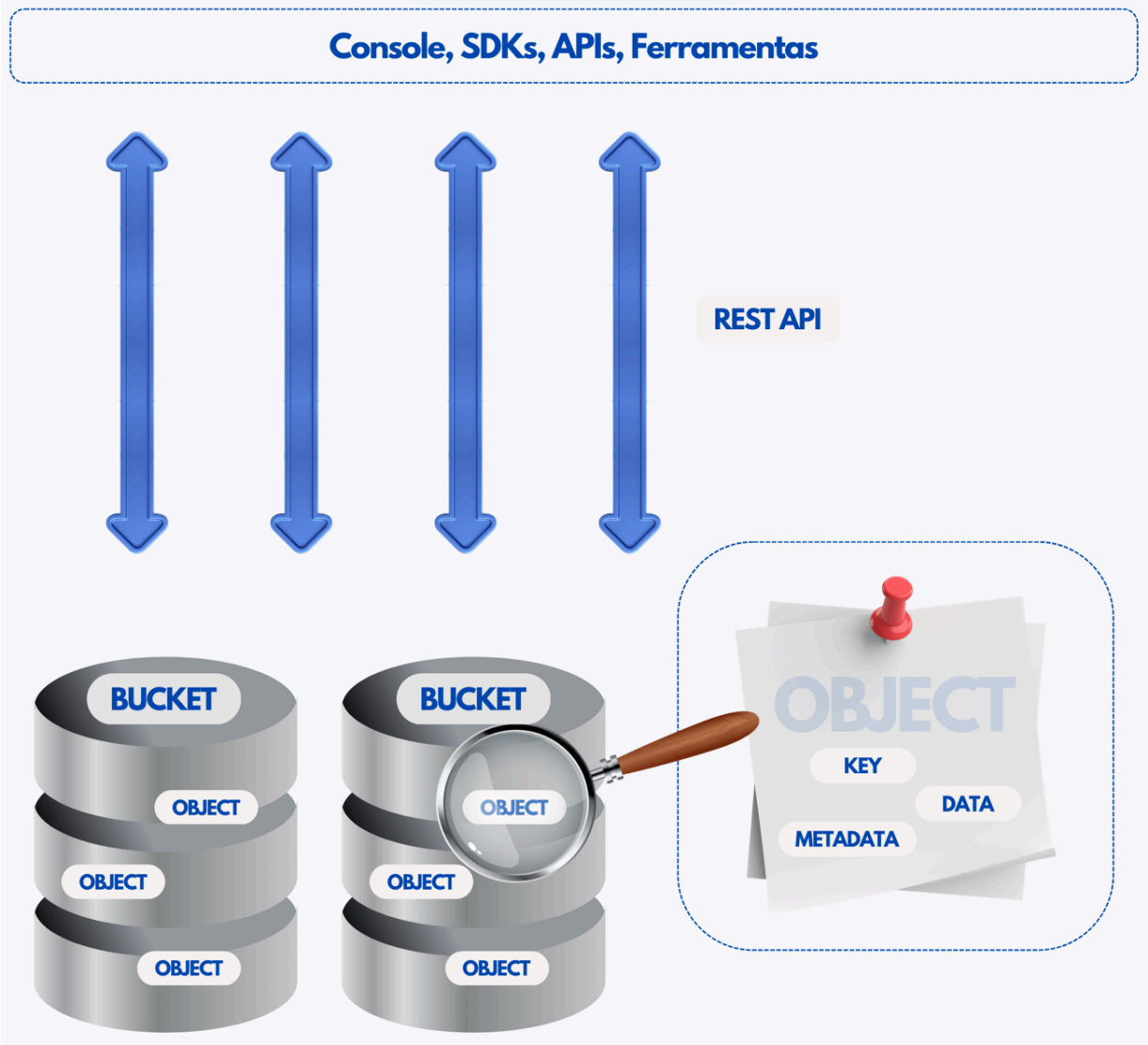


Figura 6.5: Object Storage Service - OBS, da Huawei Cloud

Um recurso relevante que merece menção nesse tipo de serviço de armazenamento é a capacidade de versionamento dos objetos. Assim como sistemas de versionamento utilizados para desenvolvimento de *softwares* ou criação de pilhas de infraestrutura como código, é possível tirar proveito dos benefícios oferecidos pelo sistema ao trabalhar com armazenamento baseado em objetos. Ao contrário de arquivos tradicionais, um objeto não pode ser editado. Caso alguma modificação seja necessária, é preciso que o usuário faça o *upload* do arquivo modificado que, ao ser gravado no *bucket*, gerará uma nova versão do objeto correspondente, armazenado

anteriormente. Por padrão, a versão mais recente de um objeto é a exibida ao usuário do serviço, mas é possível recuperar versões anteriores a qualquer tempo, caso necessário. É possível, ainda, preservar arquivos removidos pelos usuários, marcando-os com uma chave de estado que os considera como "apagados", no sentido de que não serão mais exibidos aos usuários, embora permaneçam disponíveis no *bucket* como objetos ocultos.

Vale mencionar, como diferencial dos serviços de armazenamento de objetos, a possibilidade de servir como repositório central de recursos voltado à utilização em aplicações com diferentes finalidades. Um exemplo clássico são as soluções automatizadas de criação de *backups* e *snapshots*, que normalmente são oferecidas pelos provedores de nuvem e armazenam as imagens geradas em *buckets* específicos, associados ao serviço. Uma outra área que tira grande proveito desse tipo de solução de armazenamento é a de *Big Data*. Os *buckets* podem ser criados para receber dados provenientes de várias fontes distintas de *streaming* de dados que, com o passar do tempo, se tornarão um grande e extremamente útil *data lake*, responsável por ingerir, armazenar e processar grandes volumes de dados.

As ferramentas necessárias para tratamento de dados, seleção e visualização de parâmetros já são ofertadas pela grande maioria dos provedores de nuvem, no formato de serviços, assim como soluções voltadas para a criação e treinamento de modelos de *machine learning*, que utilizarão as informações extraídas do *data lake* como matéria prima.

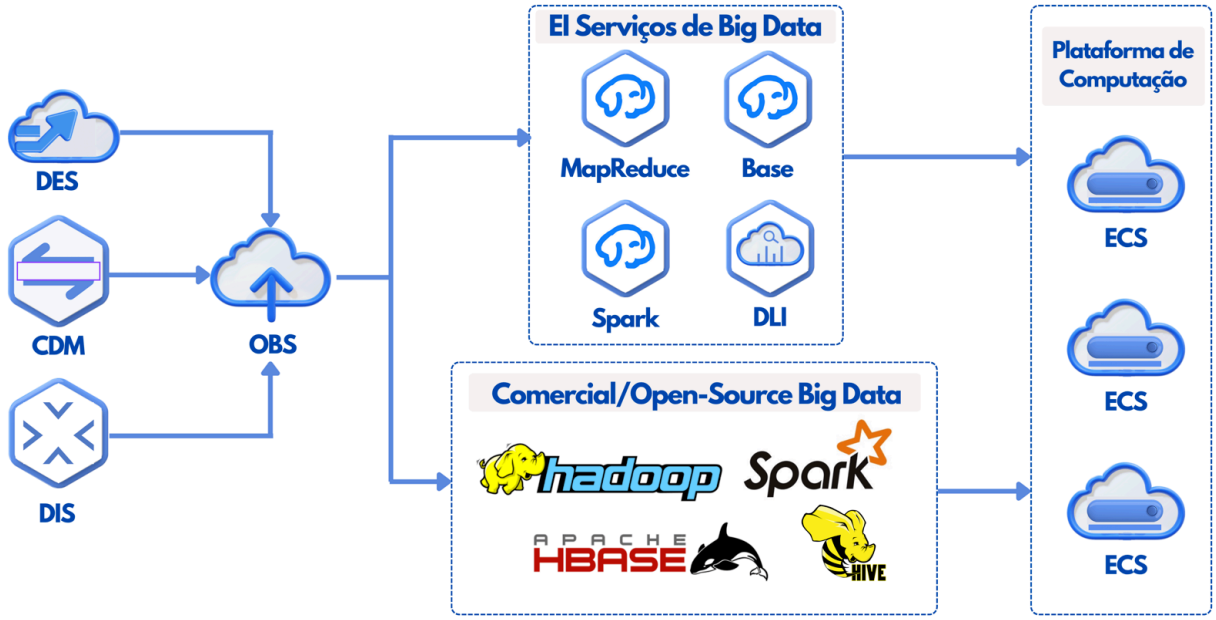


Figura 6.6: Object Storage Service - OBS, da Huawei Cloud

Outra funcionalidade muito útil que pode ser desempenhada pelos *buckets* é a possibilidade de atuar como um servidor *web*, capaz de renderizar conteúdo estático, sem a necessidade de qualquer tipo de configuração ou serviço de computação adicional. Para tanto, é necessário que o usuário habilite essa funcionalidade no *bucket* desejado, faça o *upload* dos objetos referentes ao conteúdo que se deseja disponibilizar e, por fim, modifique a configuração de visibilidade do conteúdo do *bucket* de privada para pública. Uma recomendação importante, nessa etapa, consiste na especial atenção à eventual presença de conteúdos sensíveis no *bucket*. Uma vez verificada a ausência de informações reservadas, pode-se, então, habilitar o acesso público em modo somente leitura. Vale lembrar que a renderização de conteúdos dinâmicos não é suportada pelo recurso em questão, devendo essa atribuição ficar a cargo dos serviços de computação tradicionais disponibilizados pelo provedor.

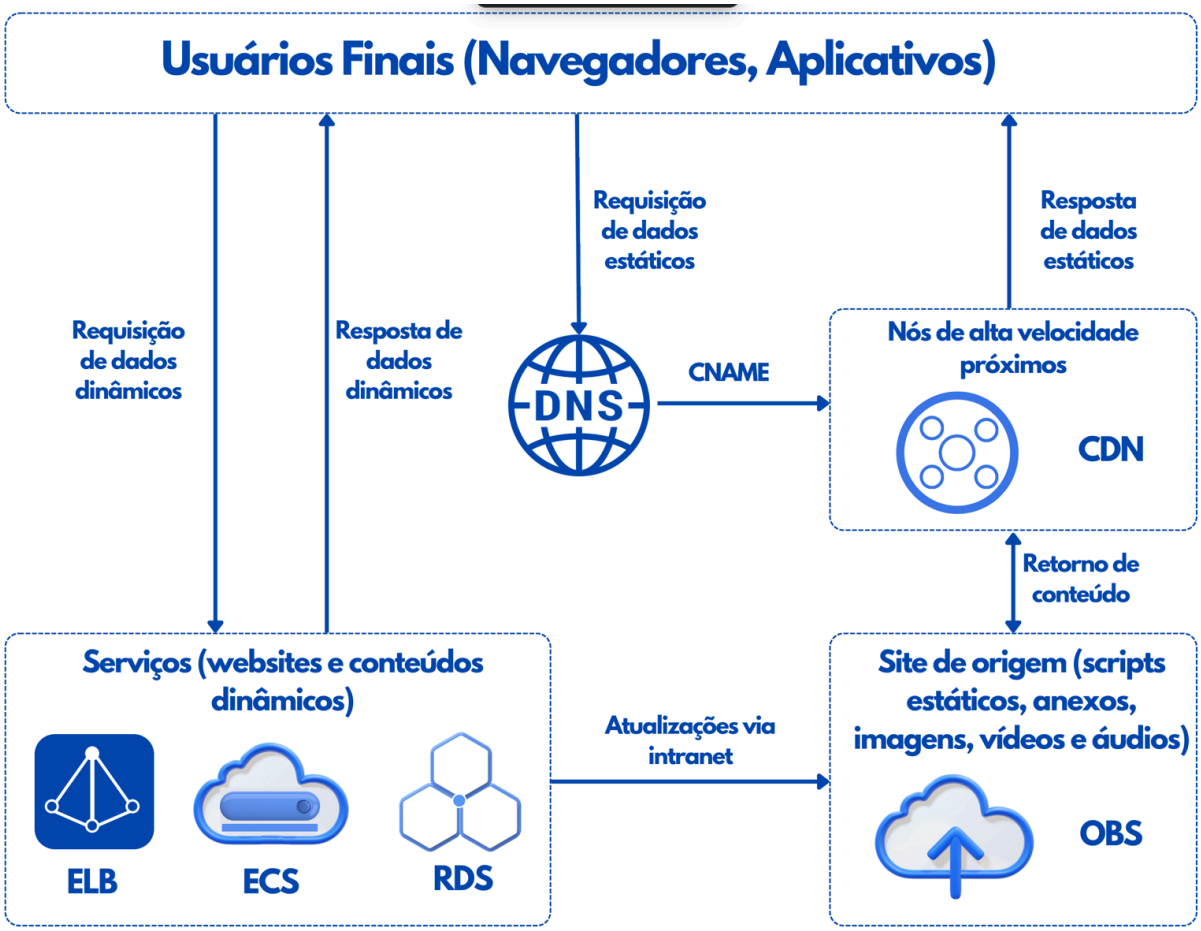


Figura 6.7: Object Storage Service - OBS, da Huawei Cloud

Classes de armazenamento

A gama de possibilidades em que um serviço de armazenamento de objetos, como o *OBS*, pode ser útil é extremamente ampla. Evidentemente, para cada possível demanda, existem particularidades, prazos de execução, prioridades e orçamentos distintos. Como lidar com tantas exceções dentro de um mesmo serviço sem torná-lo excessivamente complexo?

A resposta para essa pergunta está na criação de um recurso denominado "classes de armazenamento". Trata-se de um conjunto de funcionalidades conjugadas em diferentes *"templates"* predefinidos, oferecidos como opção ao cliente dentro do serviço. A depender da classe escolhida, é possível priorizar a confiabilidade,

excelência operacional, segurança ou a otimização de custos, por exemplo. É possível, ainda, tirar proveito de todas essas características definindo o ciclo de vida dos dados armazenados, caso as regras de negócio utilizadas na organização permitam.

A grande maioria dos provedores de nuvem pública trabalha com esse tipo de recurso. No *OBS*, em particular, existem as seguintes classes disponíveis:

- **Padrão:** a classe de armazenamento padrão, como o próprio nome sugere, é a opção utilizada pelo serviço caso o usuário não realize nenhuma personalização. Todos os dados armazenados nos *buckets* utilizando esse perfil tiram proveito de uma estratégia de armazenamento baseada na utilização simultânea de múltiplas zonas de disponibilidade, que proporciona uma durabilidade dos dados na casa de "doze noves" (99,9999999999%) e uma disponibilidade de 99,995%. Recomenda-se a utilização da classe padrão em soluções que façam uso de arquivos acessados frequentemente, pois o seu custo por *gigabyte* armazenado é o maior dentre as classes disponíveis;

- **Acesso infrequente:** possui características muito semelhantes às encontradas na classe padrão, inclusive no que se refere à estratégia de armazenamento baseada na utilização simultânea de múltiplas zonas de disponibilidade. No entanto, sua utilização é indicada em projetos em que os dados são acessados com baixa frequência. Quando comparada à classe padrão, a de acesso infrequente tem um custo de armazenamento ligeiramente mais baixo; no entanto, caso os arquivos armazenados sejam usados com frequência, essa opção se tornará menos viável economicamente, uma vez que as operações de leitura dos objetos também são precificadas nessa classe;

- **Archive:** identificada como a classe de serviço voltada ao arquivamento de dados de longa duração, que raramente serão

acessados, como *backups*, por exemplo. Possui uma tarifação de armazenamento significativamente menor do que as demais classes; no entanto, para que os dados possam ser recuperados, existe uma tarifa adicional por *gigabyte* restaurado. Vale lembrar, ainda, que o tempo de recuperação dos dados pode variar de alguns minutos até algumas horas, tendo em vista que os provedores tendem a oferecer opções com algumas variações dentro de uma mesma classe, com vistas especialmente à economicidade, o que certamente impactará o tempo de recuperação das informações desejadas.

Conforme dito anteriormente, é possível ainda definir diferentes classes a um objeto durante o seu ciclo de vida, possibilitando que um mesmo dado, embora naturalmente tenha diferentes relevâncias dentro de um determinado período de tempo, tenha sua precificação ajustada à medida que sua importância diminui. Para tanto, é necessário alterar a classe de um objeto diretamente ou movê-lo para um *bucket* que esteja pré-configurado com a classe desejada.

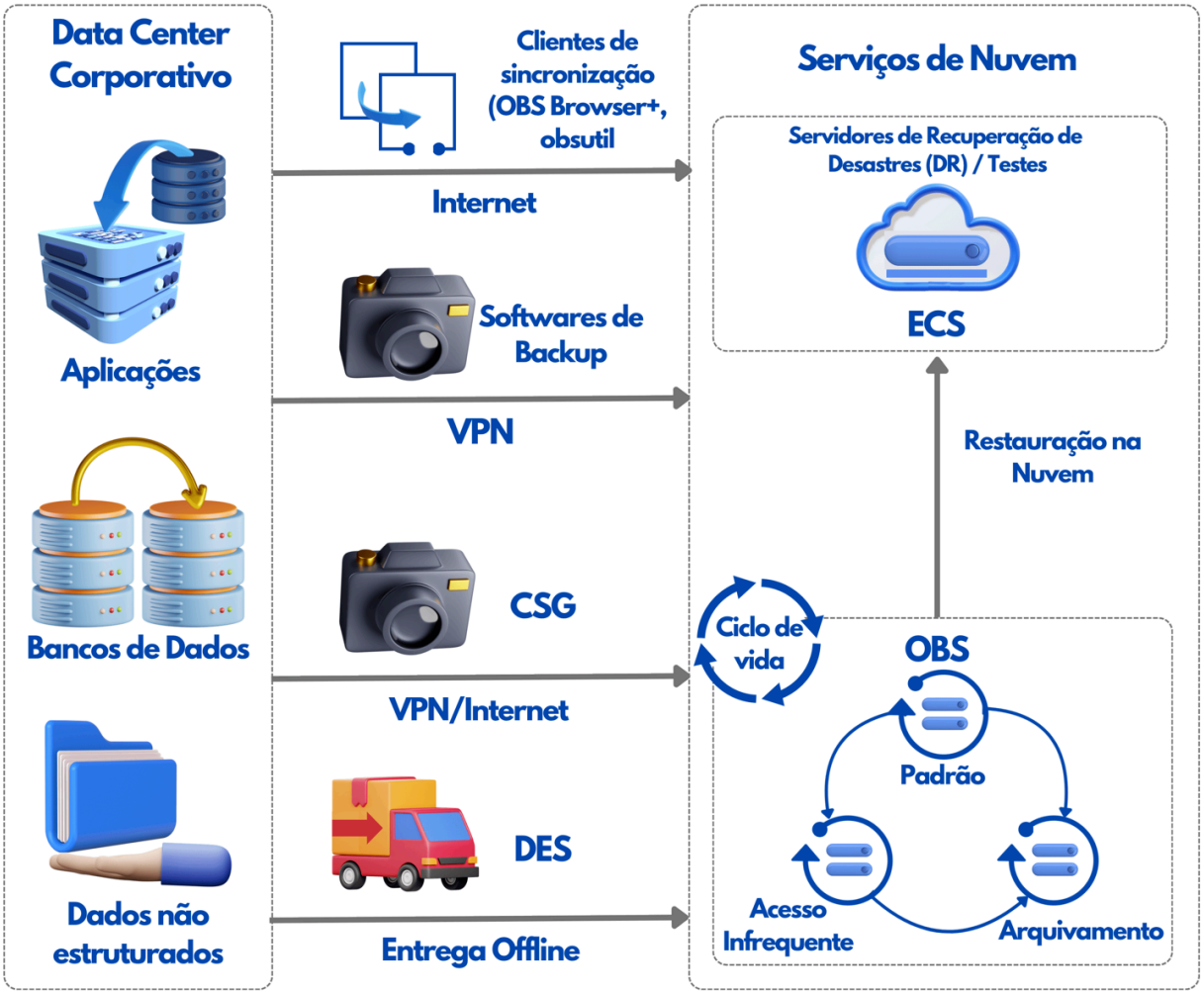


Figura 6.8: Object Storage Service - OBS, da Huawei Cloud

6.4 Serviços de migração de dados em massa

O processo de migração de dados de uma organização que utiliza infraestrutura de *datacenter* próprio para ambiente de nuvem pode ser simples ou complexo, a depender da quantidade de informações disponíveis. Em geral, o modo de transferência de dados normalmente utilizado nesse tipo de demanda consiste no envio de arquivos via internet, por túneis *HTTPS* encriptados. Na prática, os dados são copiados diretamente da origem para o repositório de

dados no destino, nesse caso, disponibilizado pelo provedor. No entanto, a depender do volume de dados a ser transferido, ou ainda em função de outros fatores, envolvendo conformidade, por exemplo, a utilização desse tipo de estratégia torna-se inviável, o que requer o uso de soluções especializadas voltadas a esse fim.

A *Huawei Cloud* contém em seu portfólio de produtos o *Data Express Service (DES)*, serviço que disponibiliza dispositivos físicos ao cliente com o objetivo de atender demandas de transferência de dados em grande escala, na casa de *petabytes*. Para tanto, oferece duas modalidades de transferência: uma baseada em discos e outra em *Teleport*. De acordo com as melhores práticas disponibilizadas pelo provedor de serviço, o seguinte padrão de uso é recomendado:

- Para volumes de dados inferiores a 30 TB: transferência baseada em disco;
- Para 30 TB ou mais: transferência baseada no *Teleport*;
- Para *petabytes* de dados: transferência com *Teleports* em cascata.

O *Teleport* é um dispositivo *plug-and-play* de armazenamento personalizado de alto desempenho, usado para transmissão massiva de dados via portas de 10GE. Conta com um sistema integrado de armazenamento empresarial da *Huawei* e suporta até 60 TB em um único dispositivo físico, podendo atender a demandas na casa de *petabytes* quando utilizado em cascata com outras unidades. Possui certificações que atestam sua resistência à poeira, água, vibrações e impactos. Além disso, dispõe de rastreamento via *GPS*, que incorpora uma série de restrições geográficas e temporais a título de segurança, bem como mecanismos de criptografia baseados em chaves assimétricas.

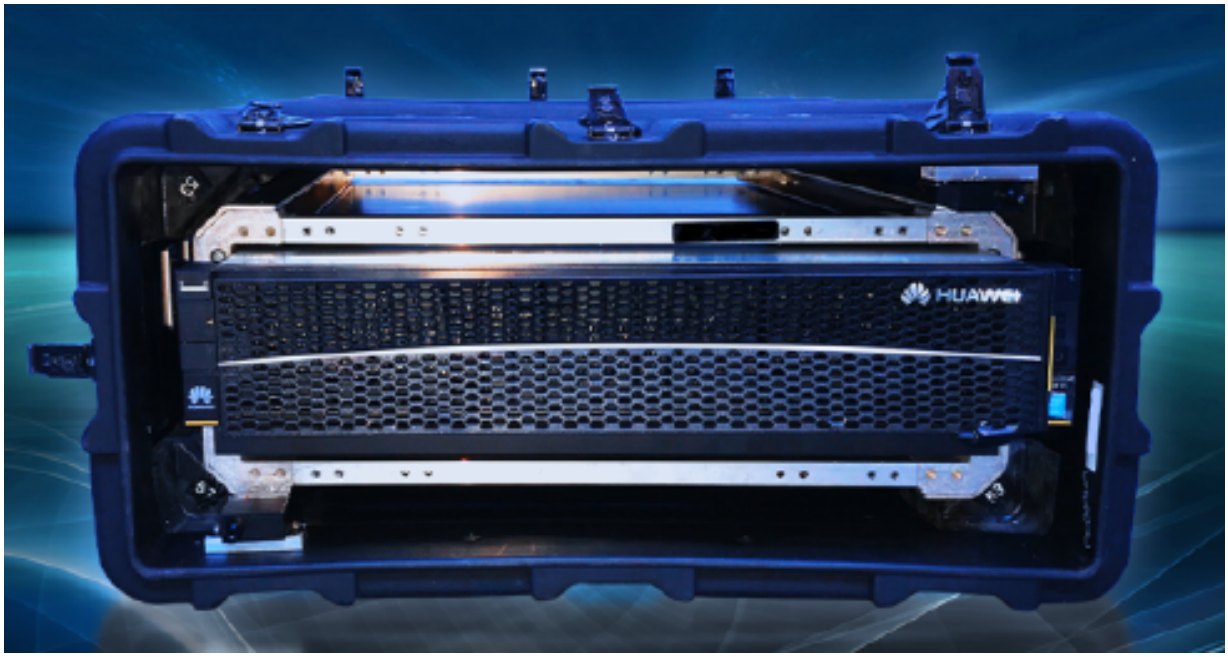


Figura 6.9: Teleport, utilizado pelo Data Express Service - DES, da Huawei Cloud

Uma vez que os discos ou o *Teleport* estejam na *Huawei Cloud*, os dados serão migrados para o *Object Storage Service (OBS)* utilizando uma política de redundância múltipla, que assegura alta durabilidade dos dados.

6.5 Considerações importantes

Os provedores de nuvem pública não cobram pelo tráfego de dados de entrada em sua infraestrutura, apenas pela quantidade de *gigabytes* armazenados. No entanto, o trânsito de saída de dados é tarifado, o que pode resultar em custos operacionais mensais acima do esperado. O mesmo ocorre com a utilização equivocada de classes de armazenamento não alinhadas ao padrão de uso da organização, ou com a não observância dos benefícios da mudança de classe de armazenamento ao longo do ciclo de vida dos dados, disponíveis nos serviços de armazenamento de objetos.

Ter ciência de fatores como esses permite ao usuário realizar estimativas periódicas de custo, simulando a influência de suas regras de negócio em aplicações reais. Isso é fundamental para que uma organização escolha, de forma assertiva, entre manter sua infraestrutura de armazenamento em ambiente *on-premises* ou optar por um conjunto de soluções oferecidas por um provedor de nuvem pública, evitando assim recorrer a soluções extremas, como a repatriação de dados, devido à falta de consideração de custos importantes no momento da decisão.

A seguir, são apresentados os serviços de armazenamento equivalentes aos abordados neste capítulo, oferecidos pelos principais provedores de nuvem pública do mercado.

Provedor	***Block Storage***	***File Storage***	***Object Storage***
Alibaba Cloud	EBS	NAS	OSS
Amazon Web Services	EBS	EFS	S3
Google Cloud Platform	Persistent Disk	Filestore	Cloud Storage
Huawei Cloud	EVS	SFS	OBS
IBM Cloud	IBM Cloud Block Storage	IBM Cloud File Storage	IBM Cloud Object Storage
Microsoft Azure	Azure Managed Disks	Arquivos do Azure	Armazenamento de Blobs do Azure
Oracle Cloud Infrastructure	OCI Block Volume	Oracle File Storage	Object Storage

Questionário online

Figura 6.10: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 6.11: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo

CAPÍTULO 7
Serviços de rede

Ao provisionarmos recursos de computação, armazenamento ou de qualquer outra natureza em um provedor de nuvem pública, é preciso ter em mente que, assim como em uma infraestrutura *on-premises*, a realização de um planejamento prévio da topologia de rede a ser utilizada é fundamental. Nesse processo, o cliente deve levar em consideração fatores como o número de endereços *IP* utilizados, a quantidade de sub-redes que devem ser criadas, a visibilidade dos recursos provisionados e quais são os elementos de conectividade necessários para que os ativos de nuvem possam ser acessados de acordo com os requisitos desejados.

7.1 Nuvens privadas virtuais

Uma das dúvidas mais recorrentes relacionadas à eventual utilização do paradigma de computação em nuvem consiste em saber como é possível provisionar, de forma segura, recursos de infraestrutura que tendem a conter informações sensíveis e privadas em um ambiente disponibilizado por um provedor de nuvem pública.

Para que possamos iniciar as considerações acerca desse tipo de questionamento, vale lembrar que o acesso administrativo necessário para o provisionamento de recursos de infraestrutura em um provedor de nuvem exige que o usuário esteja devidamente autenticado na plataforma em questão e que seu perfil possua as autorizações necessárias para a gestão dos recursos. Para tanto, é necessário que tais prerrogativas sejam concedidas previamente por meio de políticas granulares de acesso, delegadas pelo administrador da conta.

Uma vez atendidos os requisitos iniciais de autenticação e autorização, é necessário que o cliente determine onde seus recursos serão provisionados. Além da escolha da região e das zonas de disponibilidade desejadas — que representam a área geográfica e os agrupamentos de *datacenters* onde estão fisicamente localizadas as infraestruturas de computação, armazenamento e redes utilizadas para provisionamento dos ativos em nuvem pertencentes ao cliente —, é preciso definir os espaços de endereçamento a serem utilizados. Esses recursos, em especial, são disponibilizados pelas *VPCs* e suas respectivas sub-redes.

Uma *Virtual Private Cloud (VPC)*, como o próprio nome sugere, implementa uma nuvem privada virtual que, na prática, possibilita que os recursos de infraestrutura em nuvem provisionados pelo cliente estejam isolados do mundo exterior, sem conectividade com a *Internet* pública. Esse tipo de configuração é semelhante a uma *Local Area Network (LAN)*, na qual os recursos estão protegidos e se comunicam com o mundo externo apenas por meio de regras específicas.

O processo de criação de uma *VPC* consiste na escolha de um espaço de endereçamento privado definido pelo cliente. Para que os endereços possam ser utilizados, o intervalo definido inicialmente pode ser usado integralmente ou subdividido em porções menores. Em qualquer um dos casos, cada porção de endereços deve ser dimensionada de acordo com as necessidades do cliente e associada a uma ou mais sub-redes.

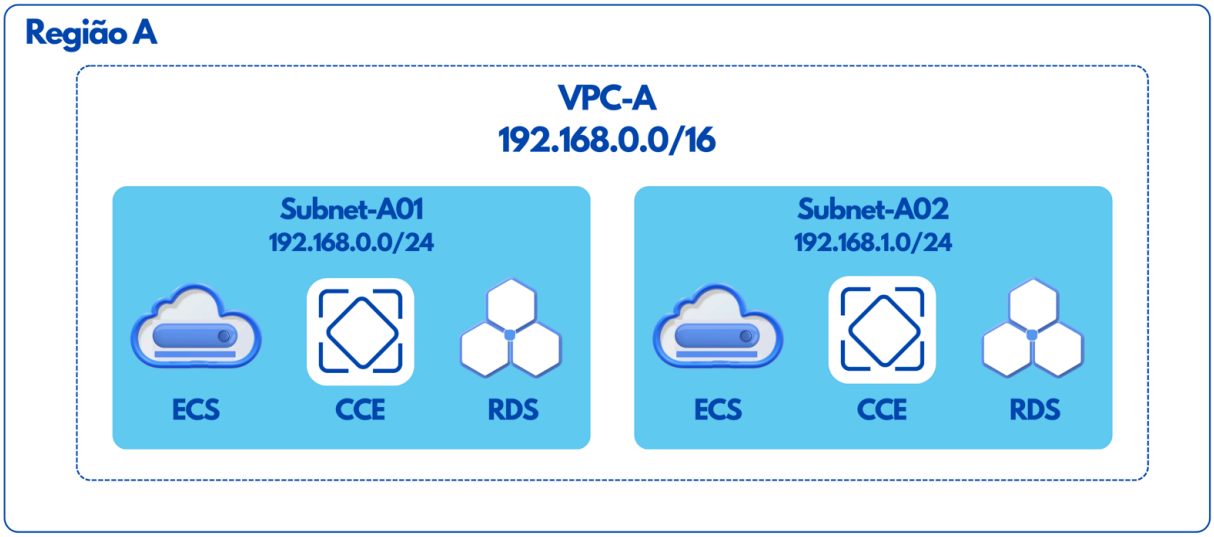


Figura 7.1: Exemplo de topologia utilizada por uma Virtual Private Cloud - VPC

Para um melhor entendimento dos conceitos envolvidos, vamos utilizar o serviço *Virtual Private Cloud (VPC)*, disponibilizado pela *Huawei Cloud*. Para localizar a ferramenta de forma rápida, basta digitar *"VPC"* na barra de filtragem localizada no topo da tela, no menu de serviços.

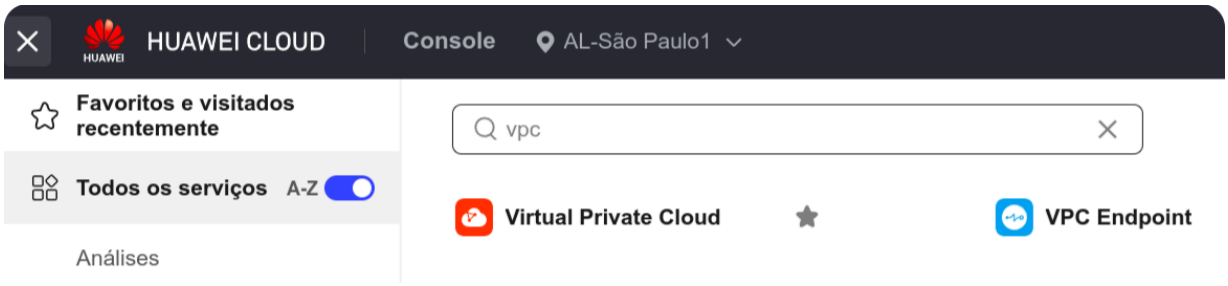


Figura 7.2: Localizando o Virtual Private Cloud - VPC no menu de serviços da Huawei Cloud

A grande maioria dos provedores de nuvem pública disponibiliza aos usuários uma *VPC* padrão em cada uma de suas regiões, de modo a facilitar o provisionamento de novos recursos, como instâncias de virtualização, contêineres, instâncias de bancos de dados gerenciados pelo provedor, dentre outros.

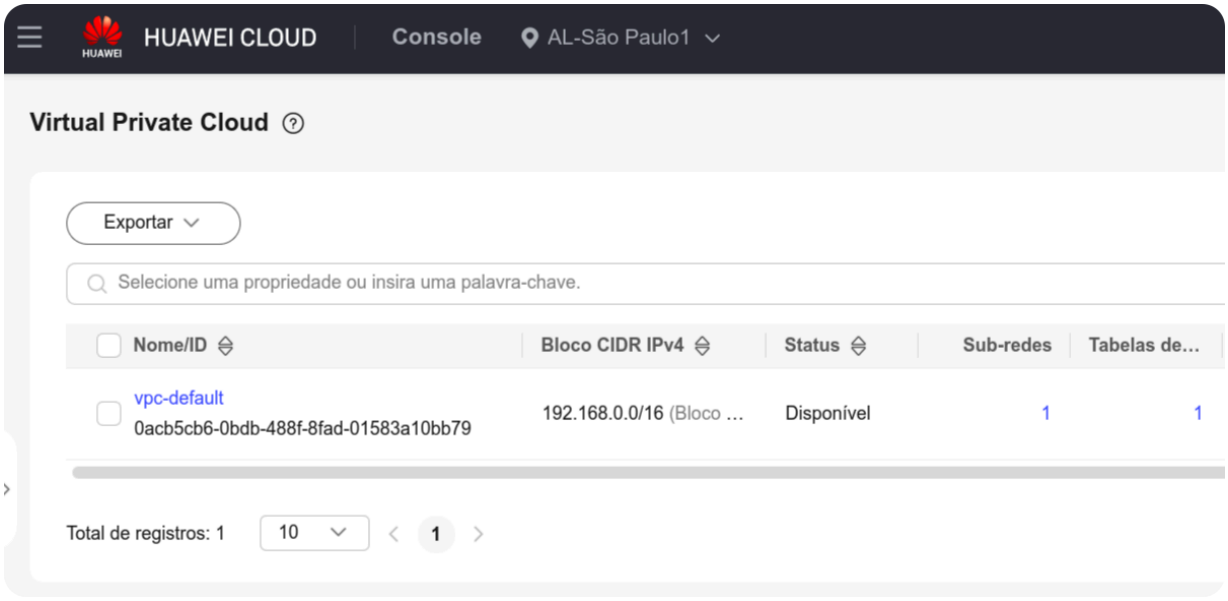


Figura 7.3: Tela inicial da interface do Virtual Private Cloud - VPC

Da mesma forma, a oferta de uma *VPC* padrão implica a necessidade da existência de ao menos uma sub-rede correspondente, para possibilitar a distribuição de endereçamento aos recursos de infraestrutura que serão provisionados internamente.

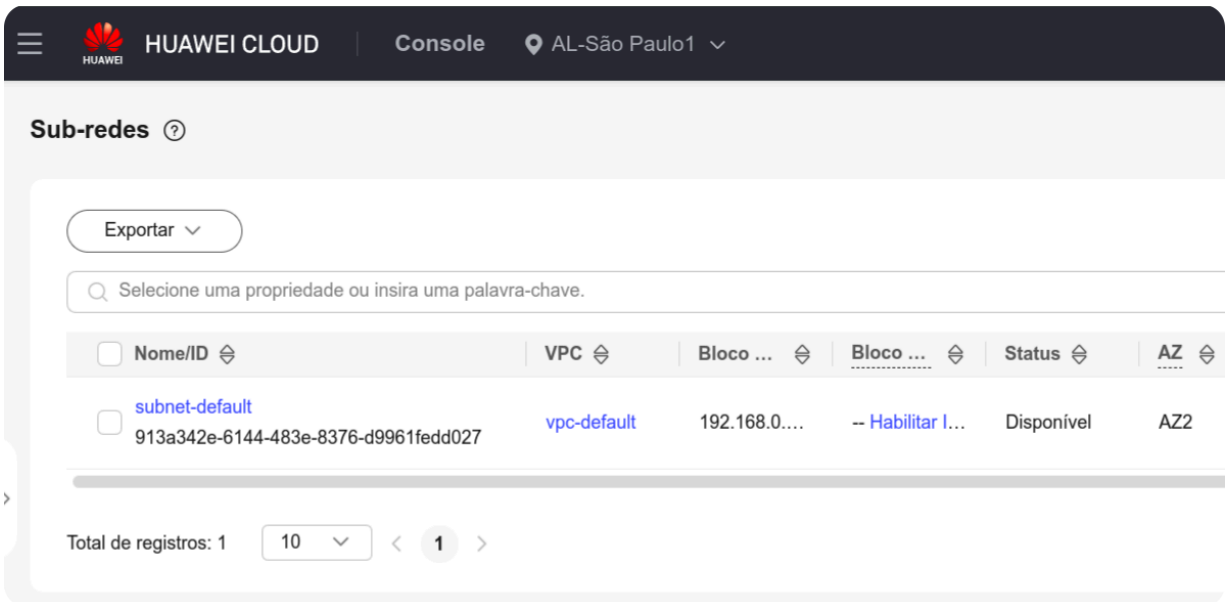


Figura 7.4: Tela inicial da interface de gerenciamento das sub-redes

Assim como em uma sub-rede tradicional, para que o tráfego de rede possa ser transmitido em uma rede privada, é necessário que exista uma tabela de roteamento associada, bem como ao menos uma rota publicada, normalmente denominada *"rota default"*. A tabela de roteamento analisa todos os pacotes de entrada e saída, encaminhando-os para o destino especificado pela rota mais aderente aos dados de cabeçalho extraídos do pacote analisado.

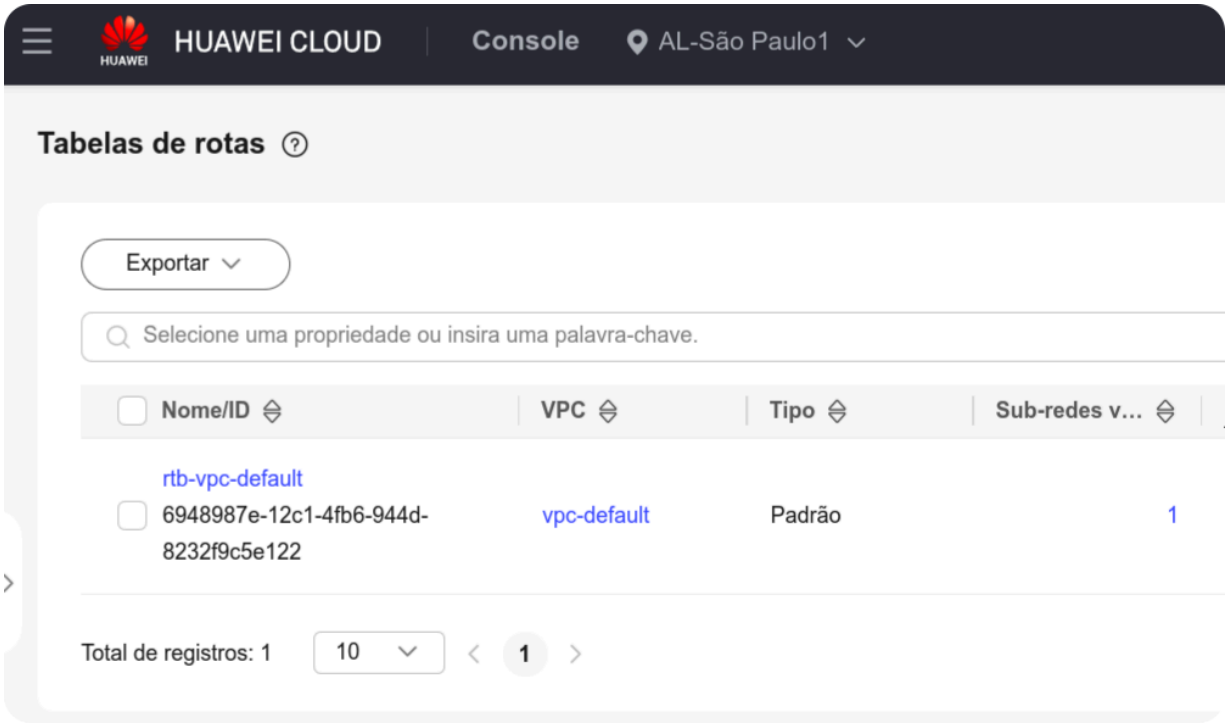


Figura 7.5: Tela inicial da interface de gerenciamento das tabelas de rotas

Em um novo projeto alocado em ambiente de nuvem, é provável que o responsável pela implementação da infraestrutura deseje isolar os recursos utilizados em uma *VPC* específica, de modo a aumentar a segurança e ter maior controle sobre as eventuais permissões concedidas aos usuários. Para atender a esse tipo de requisito, a utilização de uma *VPC default* preexistente não é recomendada.

A criação de uma *VPC* adicional pode ser facilmente realizada, bastando que o usuário clique no botão "Criar VPC", localizado no

canto superior direito da interface inicial exibida no console de gerenciamento do serviço.

O usuário deverá, então, informar a região em que deseja provisionar sua *VPC*, bem como atribuir um nome a ela. Uma boa prática sugerida nesse ponto consiste em sempre utilizar nomes significativos para descrever nossos recursos, o que facilita o entendimento da topologia utilizada à medida que os ativos de infraestrutura aumentam em quantidade. Por fim, será necessário informar qual bloco de endereçamento será utilizado na *VPC* a ser criada.

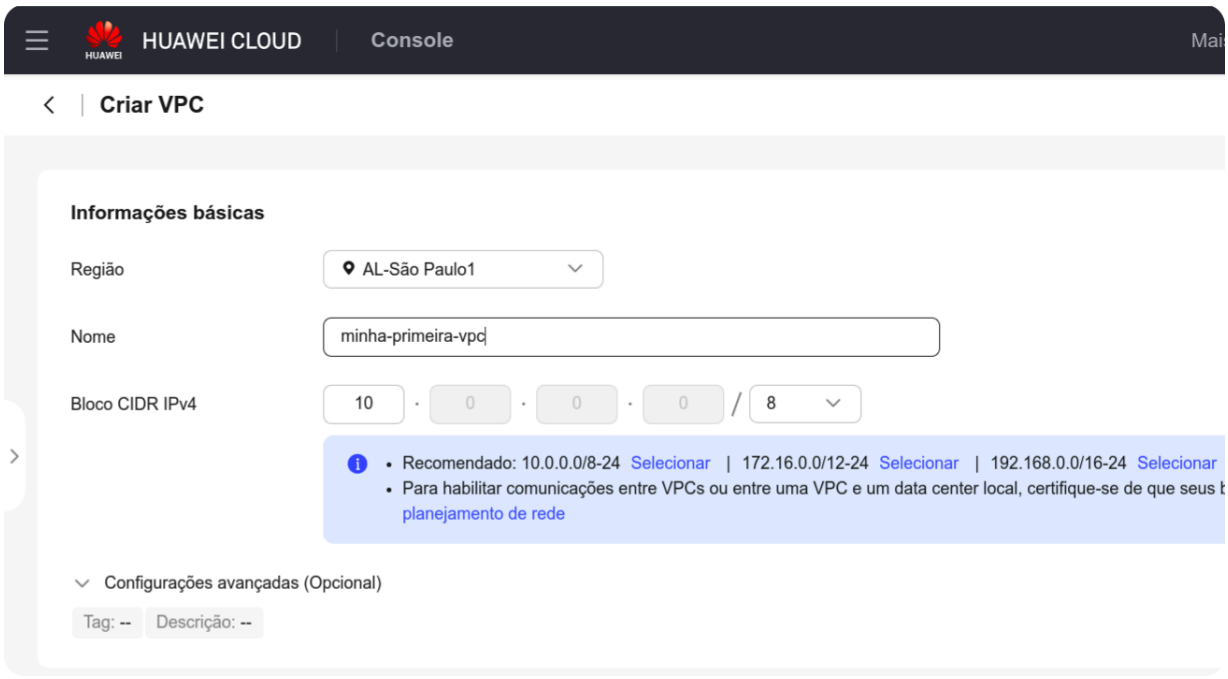


Figura 7.6: Criando uma nova VPC

No exemplo em questão, estamos utilizando uma *VPC* denominada "minha-primeira-vpc", com o prefixo 10.0.0.0/8, o que merece uma menção relevante. A *Huawei Cloud* suporta o prefixo; no entanto, caso você precise utilizar outro provedor ou esteja envolvido em projetos *multicloud*, fique atento(a), pois a grande maioria dos provedores de nuvem trabalha com *VPCs* de prefixo /16, no máximo.

Blocos de endereço CIDR suportados	Intervalo de endereços IP	Máscara de rede	Número máximo de endereços IP
10.0.0.0/8-24	10.0.0.0–10.255.255.255	8-24	2^24-2 = 16.777.214
172.16.0.0/12-24	172.16.0.0–172.31.255.255	12-24	2^20-2 = 1.048.574
192.168.0.0/16-24	192.168.0.0–192.168.255.255	16-24	2^16-2 = 65.534

Ainda no processo de criação de uma nova *VPC*, é possível criar uma sub-rede inicial para que recursos de infraestrutura possam ser, finalmente, provisionados. Para tanto, utilizaremos o nome "minha-primeira-subrede" para identificá-la e vamos alocá-la na segunda zona de disponibilidade (AZ2) da região. Vamos utilizar a primeira rede da *VPC*, com um prefixo /24. Vale lembrar que, por padrão, o endereçamento *IPv6* não está habilitado; no entanto, está disponível para uso, bastando habilitar o recurso, conforme indicado na imagem a seguir.

Algumas configurações adicionais relativas a nomes de domínio, servidores *NTP*, dentre outros recursos, também estão disponíveis, mas são opcionais. Ao clicar no botão "Criar agora", todos os parâmetros abordados serão utilizados para a criação da *VPC* e da sub-rede em questão.

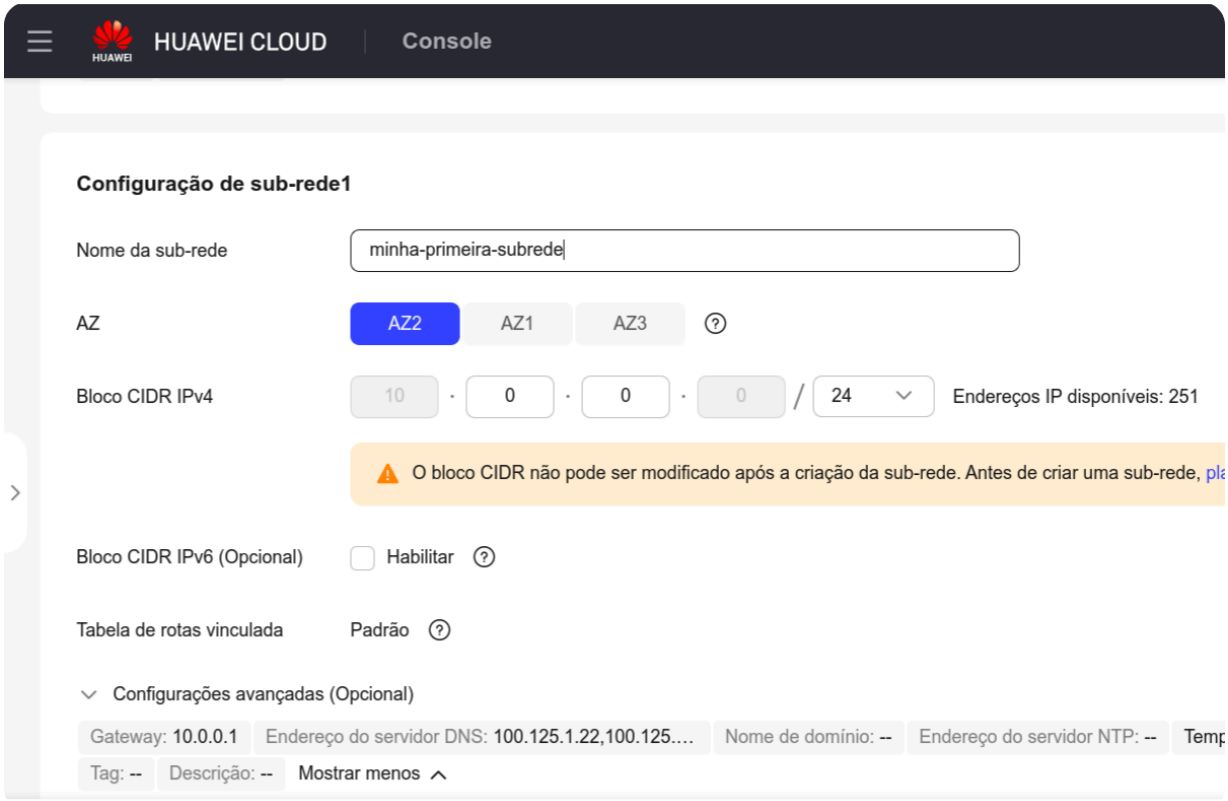


Figura 7.7: Criando uma nova sub-rede durante o processo de criação da VPC

Uma vez concluído o processo, é possível perceber que a *VPC* padrão continua disponível, juntamente do novo recurso denominado "minha-primeira-vpc". É importante observar que os prefixos utilizados em cada uma das *VPCs* disponíveis em uma região devem ser diferentes, de modo a evitar a sobreposição dos endereços. Vale lembrar também que, por definição de *design*, inicialmente não existe comunicação entre os recursos provisionados em diferentes *VPCs*.

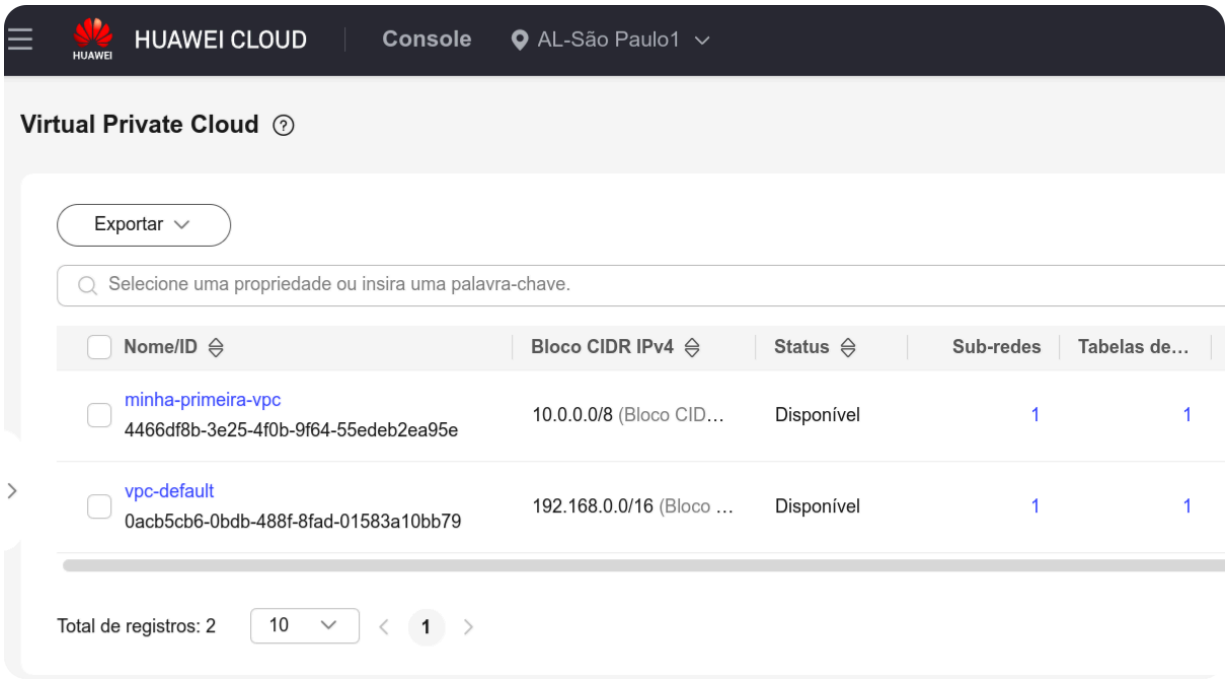


Figura 7.8: Tela atualizada da interface do Virtual Private Cloud - VPC após criação dos novos recursos

Da mesma maneira, a sub-rede denominada "minha-primeira-subrede" divide espaço com a sub-rede padrão previamente existente.

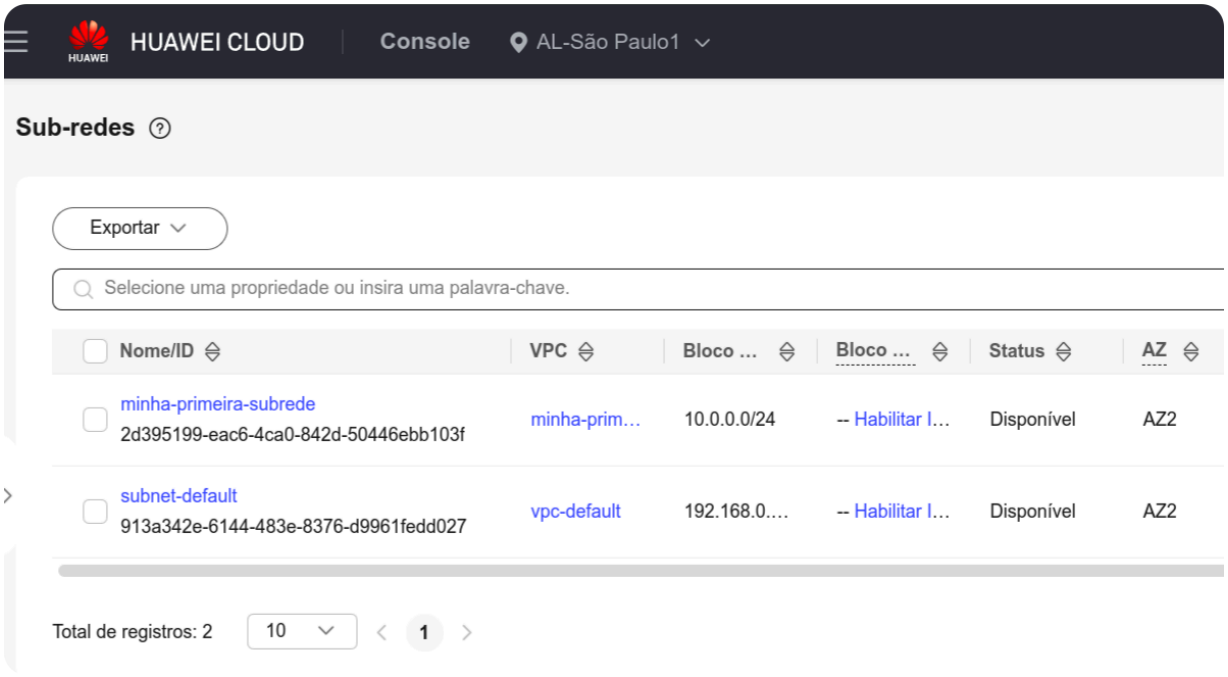


Figura 7.9: Tela atualizada da interface de gerenciamento das sub-redes após criação dos novos recursos

Naturalmente, a tabela de roteamento referente à sub-rede em questão também segue o mesmo padrão, dividindo espaço com a tabela padrão previamente existente.

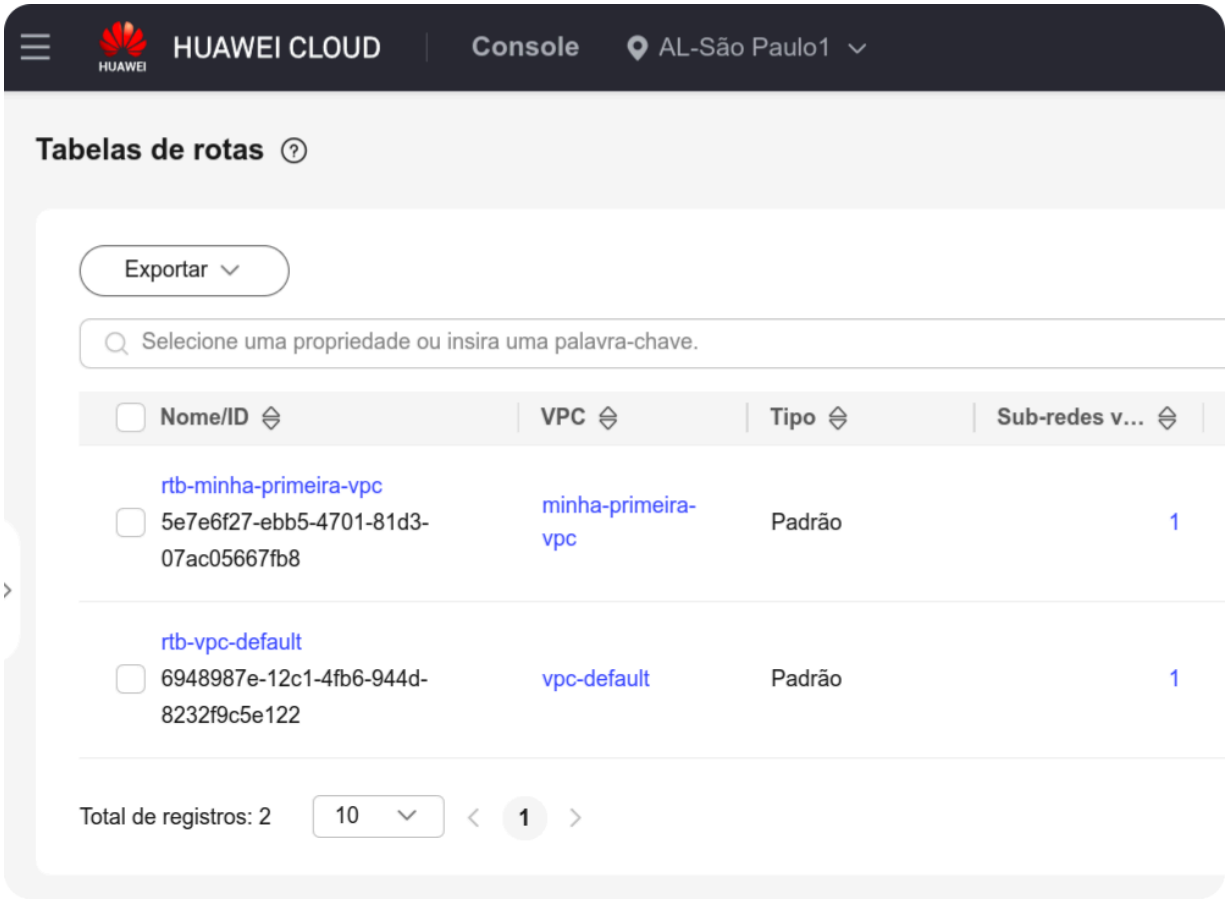


Figura 7.10: Tela atualizada da interface de gerenciamento das tabelas de rotas após criação dos novos recursos

Vale lembrar que a criação de recursos como *VPCs* e sub-redes é limitada a uma quantidade predeterminada pelo provedor de nuvem. No caso da *Huawei Cloud*, esse número é restrito a 5 *VPCs* e 100 sub-redes por conta. Tais limitações, denominadas *"soft limits"*, são flexíveis, bastando que o usuário abra um *ticket* de serviço para solicitar o eventual aumento dos quantitativos estabelecidos a princípio pelo provedor, caso considere necessário.

7.2 Tipos de conectividade

A principal característica das *VPCs* é isolar todo o tráfego de rede gerado internamente pela infraestrutura pertencente a cada uma das contas de usuários vinculadas ao provedor de nuvem pública. Dessa forma, o trânsito de pacotes está restrito somente às rotas internas que estejam ao alcance das tabelas de roteamento pertencentes à *VPC* e suas respectivas sub-redes.

No entanto, a grande maioria dos fluxos de trabalho provisionados em ambiente de nuvem requer que os servidores em operação estejam aptos a, no mínimo, realizar o *download* de aplicativos e *patches* de segurança, os quais, em geral, são aplicados no sistema operacional, nos serviços disponibilizados pelas instâncias de virtualização via rede, bem como nas demais aplicações que, porventura, estejam sujeitas a *bugs* e vulnerabilidades de *software*.

Em face do exposto, é necessário que o usuário responsável pela administração dos recursos de rede da organização adicione serviços de conectividade à topologia da conta associada ao provedor de nuvem, de modo a disponibilizar novos recursos que deverão atuar como *gateways* para a infraestrutura provisionada nas sub-redes pertencentes à *VPC* em questão.

Existem diferentes serviços dessa natureza disponibilizados pelo provedor de nuvem. A eventual utilização de uma ou outra solução deve ser pautada exclusivamente nas necessidades de negócio do cliente, que, em geral, são específicas e bem definidas.

Internet Gateway e NAT Gateway

A grande maioria dos provedores de nuvem permite o uso de modificadores de visibilidade durante o processo de criação de uma sub-rede. Geralmente, é possível classificar o recurso como público ou privado.

Ao criarmos uma sub-rede pública, devemos ter em mente que cada um dos eventuais recursos de computação ali provisionados, assim

como todo ativo de nuvem, em um primeiro momento receberá um endereço *IP* privado, utilizado para comunicação interna com os recursos de infraestrutura já provisionados. Em seguida, em função de sua característica de concentrar recursos que deverão possuir ampla visibilidade externa, receberá um endereço *IP* público, necessário para o acesso bidirecional à *Internet*. Por padrão, ambos os endereços são recebidos dinamicamente via *DHCP*, consumindo um *pool* de recursos de endereçamento disponibilizados pelo provedor — no caso do endereço público — e do respectivo intervalo de endereços atribuído à sub-rede durante seu processo de criação — no caso do endereço privado.

Ainda que os dispositivos em que se deseja obter visibilidade externa ao provedor tenham sido endereçados corretamente, é necessário que eles possuam uma rota para que possam alcançar a *Internet* e vice-versa. Para tanto, a maioria dos provedores de nuvem disponibiliza aos clientes um serviço de conectividade que, em geral, recebe o nome de *Internet Gateway* e deve ser associado diretamente à *VPC* à que se deseja dar visibilidade.

Uma vez adicionado o recurso, o usuário deve incluir uma rota padrão, representada pela notação 0.0.0.0/0, na tabela de roteamento da sub-rede pública. Dessa forma, todo o tráfego de destino que não corresponda aos endereços internos atendidos pela *VPC* e suas respectivas sub-redes será direcionado diretamente para o *Internet Gateway*. A existência prévia de *VPCs* padrão, bem como de suas respectivas *sub-redes* e tabelas de roteamento em cada região atendida pelo provedor, visa justamente poupar o cliente da complexidade de criar manualmente todos esses recursos.

Existem, ainda, cargas de trabalho em que o acesso à *Internet* é desejado e necessário, mas de forma unidirecional. Para suprir esse tipo de necessidade, os provedores de nuvem oferecem um serviço de conectividade denominado *NAT Gateway*.

Nessa modalidade, recursos de computação identificados exclusivamente via utilização de endereços *IP* privados acessam a

Internet por meio de um endereço *IP* público compartilhado, associado a uma das interfaces de rede presentes no *NAT Gateway*, que, por sua vez, também possui um endereço *IP* privado para comunicação com a infraestrutura provisionada no interior da *VPC*.

Essa metodologia de acesso utiliza a mesma técnica das conexões domésticas à *Internet*, em que o provedor de acesso entrega um endereço *IP* público dinamicamente alocado à interface *WAN* do modem, que possibilita o acesso de todos os dispositivos existentes na rede local via interface *LAN*, sem que estes estejam visíveis diretamente na rede mundial de computadores, preservando a segurança sem comprometer a conectividade.

Elastic IP (EIP)

No momento em que um cliente utiliza um *IP* público em seus recursos de infraestrutura provisionados em suas sub-redes também públicas, o endereço entregue ao dispositivo final é um *IP* roteável na *Internet*, porém obtido de forma dinâmica.

Em termos práticos, isso significa que o endereço em questão não é estático, ou seja, pode mudar conforme a ocorrência de determinado evento. Um dos principais gatilhos para esse tipo de mudança está associado ao momento em que uma instância de virtualização tem seu ciclo de vida modificado.

Uma vez em estado de "execução" e com um endereço público atribuído, quando a instância recebe um sinal para assumir o estado de "parar", o endereço em questão é liberado. Mesmo que um novo sinal — desta vez para mudar ao estado de “iniciar” — seja recebido pela instância, será entregue um novo endereço dinâmico do *pool* de endereços IP públicos disponibilizado pelo provedor.

Por outro lado, caso a instância receba um sinal indicando o estado de "reiniciar", o endereço público é mantido, não havendo alteração.

Para compreendermos o processo, é importante levarmos em conta que, ao reiniciar uma instância de virtualização, seus recursos de

provisionamento seguem alocados em memória, o que garante que o nó físico de processamento pertencente à infraestrutura do provedor de nuvem responsável pelo seu gerenciamento continue a lidar com a mesma função.

Quando a instância de virtualização é parada, não há como o *hypervisor* responsável pela gerência do ambiente de virtualização prever quando o recurso será novamente ativado. No momento em que a instância for novamente iniciada, é muito provável que sua execução esteja sob responsabilidade de um nó físico de processamento diferente do anterior, eventualmente alocado em um outro *rack* do *datacenter*, o que invariavelmente impactará nos endereços *IP* públicos a que tem acesso — daí a necessidade de atualização do endereçamento.

Caso o cliente tenha a necessidade de um endereço *IP* público estático, é necessário utilizar um *Elastic IP*. Entre os recursos oferecidos pelo serviço, destacam-se a facilidade de vincular e desvincular o endereço entre diferentes instâncias ou modalidades de serviço, além da gestão customizada da largura de banda utilizada.

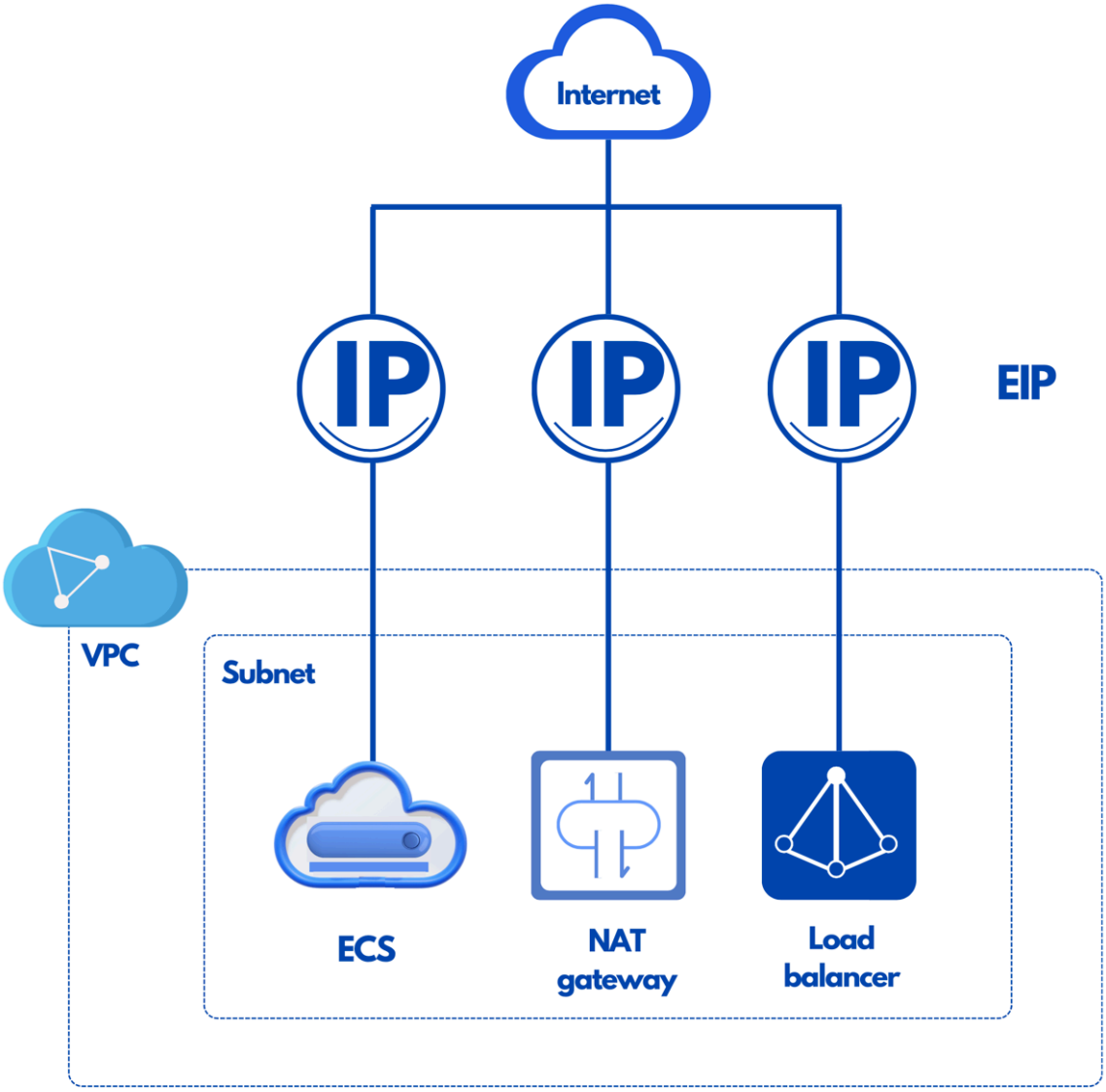


Figura 7.11: Topologia utilizada na atribuição de endereços IP elásticos - EIPs

IP Virtual

Ao provisionar recursos de computação em qualquer tipo de infraestrutura, é fundamental que o administrador implemente servidores redundantes para serviços de missão crítica. O principal benefício dessa abordagem é mitigar cenários em que exista um ponto único de falha.

Em ambientes de nuvem, a adoção de recursos de computação que utilizam estratégias de escalabilidade horizontal, em conjunto com uma solução de balanceamento de carga, é uma prática amplamente adotada. Contudo, a depender do recurso financeiro disponível para a manutenção de um projeto específico, essa abordagem pode representar um custo adicional significativo, tornando sua implementação inviável do ponto de vista econômico.

Tal afirmação baseia-se em cenários nos quais os custos envolvidos desempenham papel determinante na tomada de decisão na implantação do projeto, tendo em vista que cada uma das réplicas utilizadas nessa arquitetura permanece em funcionamento durante todo o tempo.

Por outro lado, é possível implementar um cenário paliativo que mantenha a redundância das informações e elimine a ocorrência de um ponto único de falha, por meio da estratégia de utilizar uma réplica do servidor original que deverá atuar como secundário.

Nessa abordagem, o servidor principal atende às requisições entrantes, enquanto o secundário fica em *standby*, mantendo-se em repouso e, portanto, sem gerar altos custos de computação. O servidor em repouso só entra em atividade caso a instância principal apresente algum tipo de falha de funcionamento, tomando seu lugar e passando a atender às demandas de acesso.

O recurso necessário para a implementação em questão consiste em um serviço denominado pelos provedores de nuvem de *IP Virtual*. Trata-se de um endereço de serviço que pode ser adicionado como um endereço *IP* adicional a, pelo menos, duas instâncias de destino: uma primária e uma secundária.

As instâncias utilizadas na implementação possuem um endereço privado cada, utilizado somente para fins de gerenciamento. O serviço de rede a ser oferecido na modalidade em questão deve estar previamente instalado e replicado nas duas instâncias. Vale lembrar que as configurações devem ser mapeadas para executar o

serviço utilizando o endereço secundário, no caso, o atribuído ao *"IP Virtual"*. Por definição, a instância primária responde pelo endereço em questão, que é transferido para a secundária somente em caso de falha.

Para detectar eventuais falhas entre os nós primário e secundário, as instâncias possuem conectividade ponto a ponto para realizar testes periódicos de saúde dos *hosts*. Essa metodologia de monitoração é amplamente utilizada em serviços com mesmo propósito, como o *Heartbeat*, muito comum em sistemas *Linux*. Contudo, existem abordagens mais modernas que não requerem conexão dedicada entre os nós, como a implementação do *CARP*, presente em sistemas *BSD*.

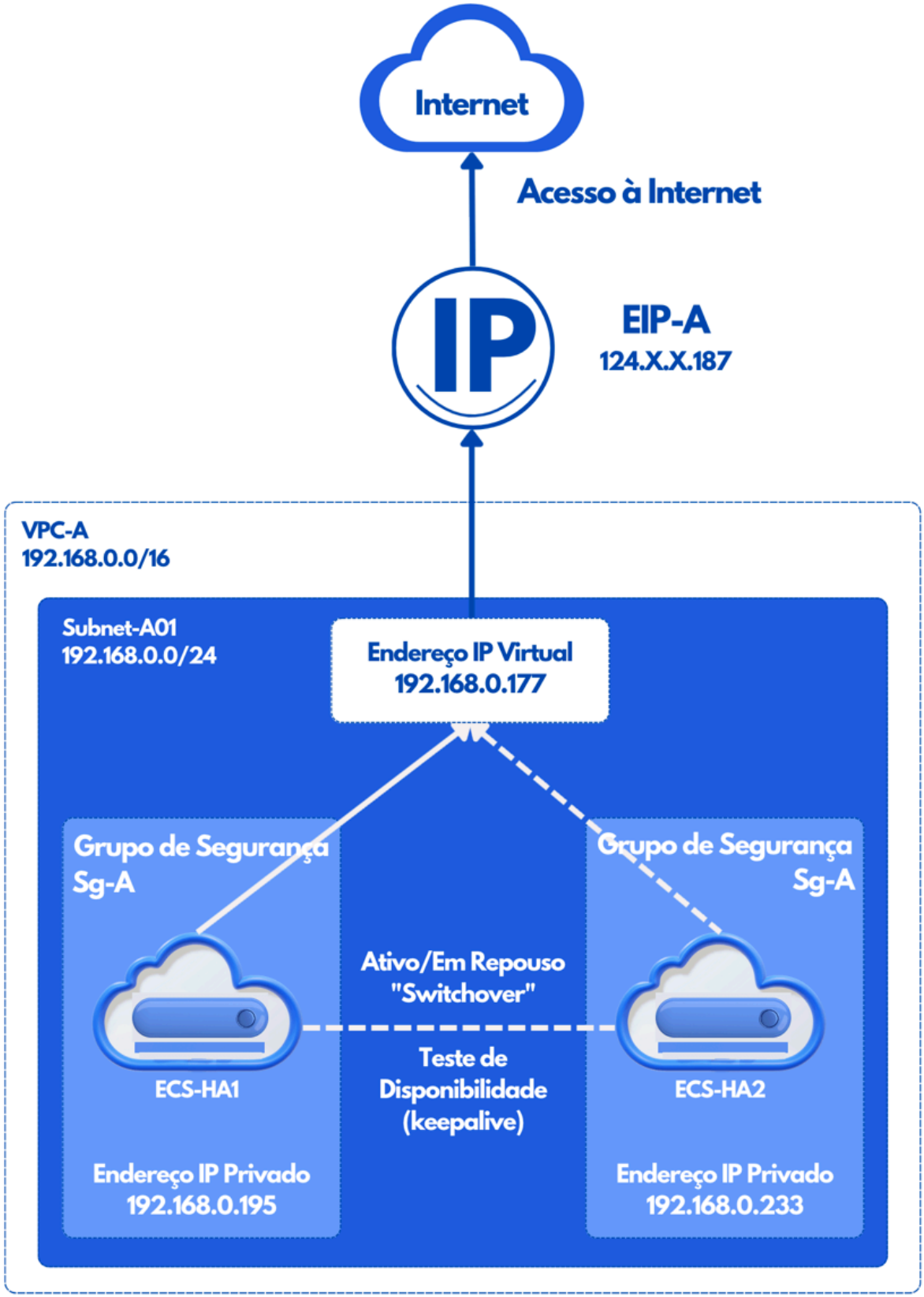

Internet
Acesso à Internet
IP
EIP-A
124.X.X.187
VPC-A
192.168.0.0/16
Subnet-A01
192.168.0.0/24
Endereço IP Virtual
192.168.0.177
Grupo de Segurança
Sg-A
Grupo de Segurança
Sg-A
Ativo/Em Repouso
"Switchover"
Teste de
Disponibilidade
(keepalive)
ECS-HA1
ECS-HA2
Endereço IP Privado
192.168.0.195
Endereço IP Privado
192.168.0.233

Figura 7.12: Topologia utilizada na atribuição de um endereço IP virtual

Virtual Private Network (VPN)

Dentre os principais tipos de implantação de computação em nuvem disponíveis na atualidade, a nuvem híbrida destaca-se por sua ampla adesão, pois possibilita ao usuário provisionar parte da infraestrutura no provedor de nuvem pública e continuar utilizando a infraestrutura legada *on-premise* preexistente, preferencialmente já tendo adotado uma solução de nuvem privada para a gestão dos recursos.

No entanto, para que as duas estruturas descritas possam caracterizar efetivamente a condição de nuvem híbrida, é necessário que os recursos em questão possam se comunicar entre si. Para tanto, a solução recomendada consiste, em um primeiro momento, na utilização de uma *Virtual Private Network (VPN)*, dada sua ampla compatibilidade e facilidade de implantação.

Em geral, os provedores de nuvem pública oferecem ao menos dois tipos de conectividade *VPN*:

- **host to site:** conecta um cliente — geralmente um *host* provisionado dentro da infraestrutura do cliente, ou ainda um dispositivo móvel, como um *notebook* — ao *gateway VPN* existente no provedor de nuvem. Uma vez estabelecida a conexão ponto a ponto entre os dois dispositivos, é criado um túnel virtual responsável por encriptar todas as informações trocadas entre o *host* de origem e a infraestrutura de destino.
- **site to site:** utiliza o mesmo princípio; no entanto, a conexão é realizada entre um *gateway VPN* localizado na infraestrutura do cliente e outro existente no provedor de nuvem. Dessa forma, em vez de conectar um único *host* à infraestrutura provisionada no provedor de nuvem, é possível conceder acesso a todos os dispositivos que, porventura, utilizem o *gateway VPN* como rota para o destino supracitado.

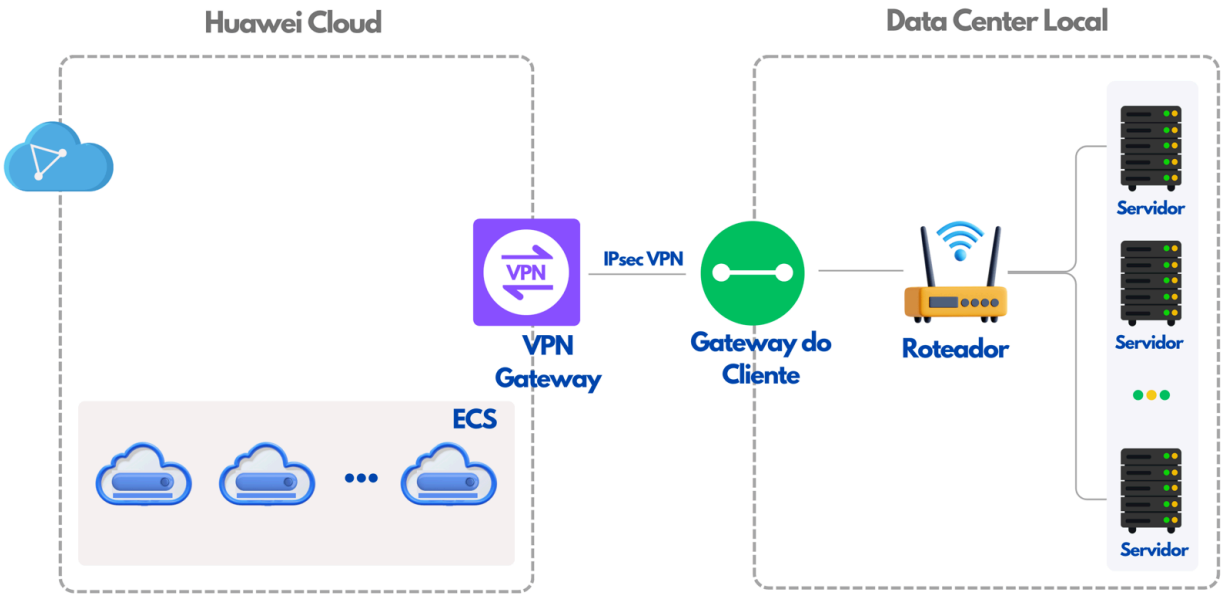


Figura 7.13: Topologia utilizada na comunicação entre a infraestrutura on-premises do cliente e a disponível no provedor de nuvem utilizando o recurso de VPN

Direct Connect

Apesar de a *VPN* representar o tipo de conectividade mais comum em casos de uso cujo objetivo principal é comunicar os ativos computacionais provisionados em uma infraestrutura de nuvem e os eventualmente preexistentes nas instalações do cliente, trata-se de uma alternativa que utiliza a *Internet* pública para o trânsito de dados. Dessa maneira, apesar de prática e acessível, a solução está sujeita aos possíveis problemas existentes em um ambiente com tais características, especialmente os relacionados à disponibilidade, estabilidade, segurança e velocidade.

Nesse sentido, considera-se uma boa prática utilizar a solução de *VPN* apenas para fins de testes de infraestrutura e homologação de sistemas que, eventualmente, serão utilizados em ambiente de nuvem híbrida.

Para a utilização cotidiana, recomenda-se, em um segundo momento, a migração para uma solução de conectividade ponto a ponto dedicada, normalmente denominada *"Direct Connect"*. Essa

solução é geralmente ofertada por uma rede de parceiros designada e homologada pelo provedor de nuvem, e disponibiliza uma conexão direta via fibra óptica com alta largura de banda entre o cliente e o provedor de nuvem, sem trânsito pela *Internet* pública, possibilitando comunicação de baixa latência e estabilidade no trânsito de dados.

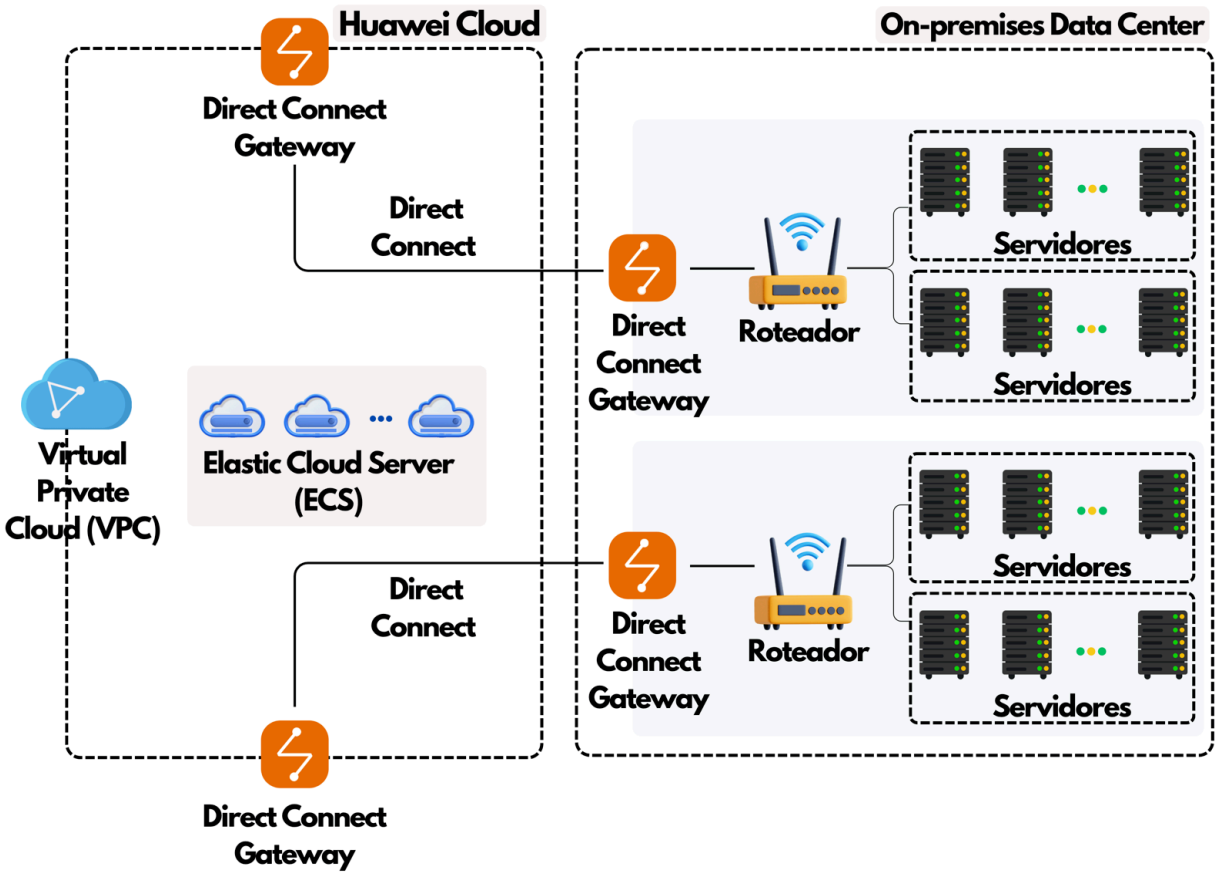


Figura 7.14: Topologia utilizada na comunicação entre a infraestrutura on-premises do cliente e a disponível no provedor de nuvem utilizando o recurso Direct Connect

Vale lembrar que, por se tratar de uma funcionalidade que, em geral, depende de uma ampla rede de parceiros, o serviço em questão pode não estar disponível para uso em muitas localidades, especialmente nas mais distantes dos grandes centros.

VPC Peering

Os provedores de nuvem pública disponibilizam *VPCs* padrão, bem como seus respectivos componentes de conectividade, sub-redes e tabelas de roteamento, em cada região onde atuam. Essa iniciativa visa tornar o provisionamento de infraestrutura mais intuitivo, tornando a experiência dos clientes — especialmente os iniciantes — mais agradável.

Da mesma maneira, usuários mais avançados ou com necessidades específicas de segregação de infraestrutura por projetos ou outros atributos podem criar suas próprias *VPCs* adicionais para atender a casos de uso específicos.

Como a finalidade de uma *VPC* é justamente isolar os recursos provisionados em sua estrutura, a princípio não existe comunicação disponível entre duas ou mais *VPCs* vizinhas. No entanto, assim como existem *gateways* especializados para necessidades de conectividades específicas, podemos recorrer a uma solução voltada justamente para esse propósito, denominada *"VPC Peering"*.

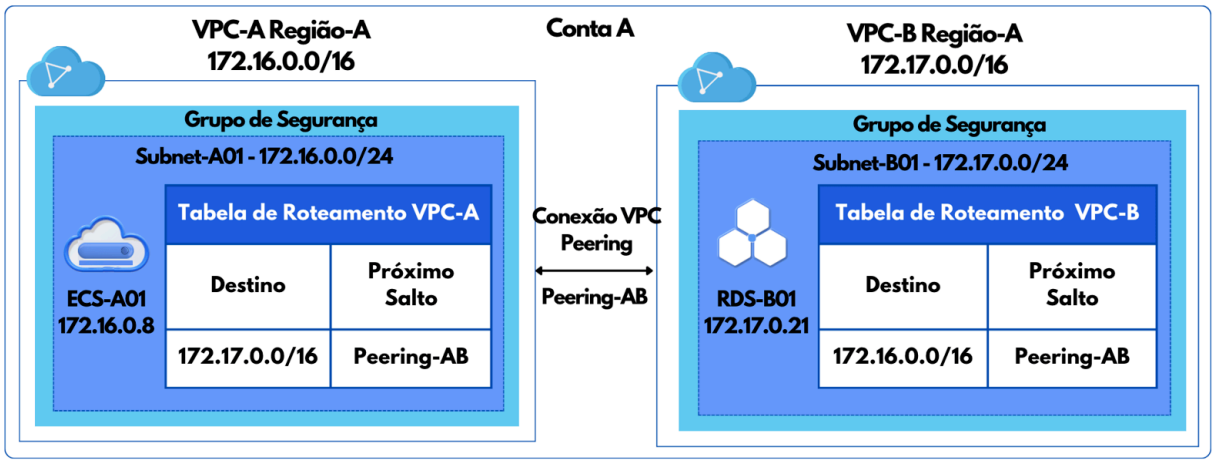


Figura 7.15: Conectando duas VPCs com VPC Peering

7.3 Considerações importantes

A utilização de diferentes serviços de conectividade em ambiente de nuvem é uma questão que deve ser dominada por profissionais da área. Os possíveis cenários de utilização, bem como a combinação entre os serviços disponíveis, são diversos e devem estar totalmente alinhados com as necessidades do cliente e do projeto. Vale lembrar que as soluções de conectividade eventualmente utilizadas devem estar alinhadas com as boas práticas de segurança adotadas em ambiente de nuvem, em especial, o princípio do privilégio mínimo — conceito que será detalhado no capítulo 10, totalmente dedicado ao tema de segurança.

Provedor	Nuvem Privada Virtual
Alibaba Cloud	VPC
Amazon Web Services	VPC
Google Cloud Platform	VPC
Huawei Cloud	VPC
IBM Cloud	IBM Cloud VPC
Microsoft Azure	Rede Virtual do Azure
Oracle Cloud Infrastructure	VCN

Questionário online

Figura 7.16: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 7.17: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo

Parte 4: Serviços gerenciados pelo provedor de nuvem

Capítulo 8
Serviços de monitoração

A oferta de serviços de rede em ambiente *on-premises* é uma prática rotineira para a grande maioria dos profissionais responsáveis pela administração da infraestrutura de uma organização. Nesse contexto, diferentes tecnologias podem ser empregadas para o provisionamento de serviços de computação, variando desde alternativas mais conservadoras — como a utilização dos tradicionais servidores dedicados — até soluções mais atuais, que se destacam pela eficiência no uso dos recursos disponíveis, como a virtualização, a orquestração de contêineres e as plataformas de nuvem privada.

Independentemente da solução escolhida, que inclusive pode ser baseada na utilização de diferentes estratégias de forma combinada, é necessário que os ativos de *hardware* e *software* sejam monitorados constantemente, de modo a antever eventuais problemas de forma preventiva ou, ainda, permitir ao gestor de infraestrutura entender fatos sequenciais que culminaram em um incidente, a fim de que seja realizada uma intervenção corretiva.

Para tanto, a grande maioria dos serviços de rede oferece um sistema de *log* próprio, que, em geral, utiliza um serviço específico para a gerência desse tipo de recurso, como o amplamente adotado *syslog-ng*.

De forma simplificada, podemos dizer que os *logs* de serviço registram eventos realizados entre duas entidades específicas e bem-definidas: os clientes e o servidor em questão. Em geral, os

eventos registrados resumem-se a atividades de acesso bem-sucedidas, armazenadas em um arquivo com nomenclatura significativa, como *"access.log"*, por exemplo. Da mesma maneira, eventos que representam justamente o oposto — ou seja, tentativas de operações que não foram executadas conforme o esperado — são registrados em um segundo arquivo, denominado *"error.log"*.

Evidentemente, as nomenclaturas mencionadas dos arquivos de *log* são usuais, mas não necessariamente mandatórias. Alguns serviços, inclusive, concentram os registros de tentativas de acesso bem e malsucedidas em um arquivo único, com o nome do *"daemon"* e a extensão *.log* ao final.

Juntamente aos eventos, estão as respectivas *timestamps*, com o dia da semana, data, hora e descrição do registro em questão. Em geral, existem diferentes níveis de criticidade de registro, que podem ser modificados a qualquer tempo pela pessoa administradora da infraestrutura, a depender de suas necessidades de levantamento de informações. Níveis mais altos, normalmente classificados como *"debug"*, tendem a registrar praticamente qualquer atividade. Em face do exposto, esse nível de registro gera grande volume de informações de armazenamento, sendo indicado somente em casos específicos ou em cenários de *troubleshooting*, nos quais é necessário solucionar problemas não previstos ou documentados.

Já o monitoramento de equipamentos é realizado por meio da utilização de protocolos específicos, como o *Simple Network Management Protocol - SNMP* ou, ainda, através de soluções de *software* amplamente utilizadas, voltados à monitoração e observabilidade, como o *Zabbix*, por exemplo.

8.1 Monitoração de infraestrutura

Tão importante quanto a iniciativa de utilizar ferramentas de observabilidade é a prática cotidiana do administrador de infraestrutura de analisar os *logs* gerados pelos ativos que estão sob sua responsabilidade. Caso essa rotina não ocorra, a organização tem em mãos somente um quantitativo de dados brutos que cresce a cada dia, mas sem utilidade real.

Em uma organização de médio porte, cumprir esse tipo de requisito já não é uma tarefa trivial. Por esse motivo, é necessário que a pessoa administradora de redes tenha em mente que a automatização de processos é um objetivo a ser perseguido sempre que possível.

Dentre as ferramentas disponibilizadas pelos provedores de nuvem pública, destacam-se as pertencentes à categoria de gerenciamento e governança. A modalidade em questão conta com serviços voltados à monitoração de recursos de infraestrutura e de serviços de rede, que atuam de forma semelhante aos utilizados em ambiente *on-premises*.

Na *Huawei Cloud*, o principal serviço utilizado para a monitoração de infraestrutura provisionada pelo cliente, denomina-se *Cloud Eye*.

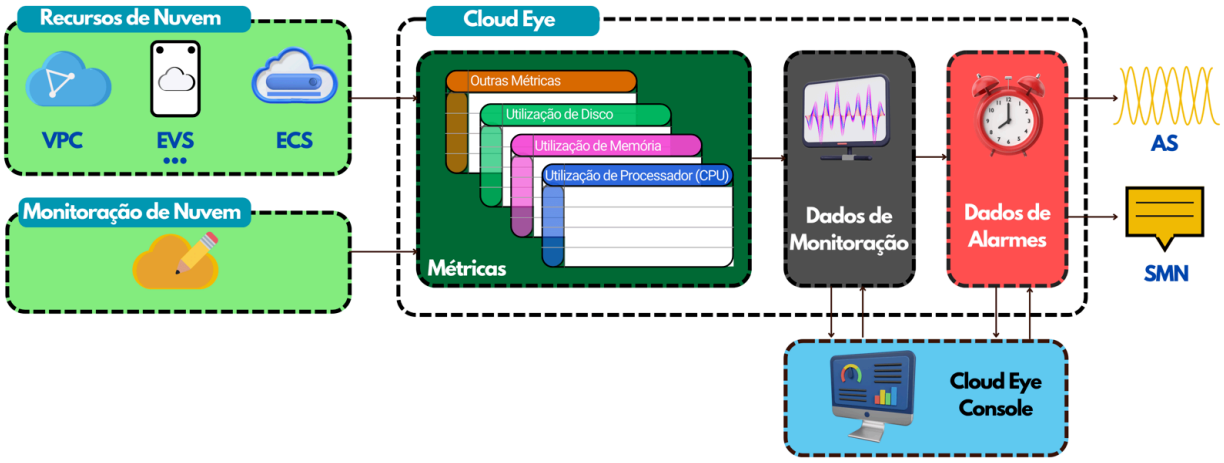


Figura 8.1: Monitoração de infraestrutura realizada pelo serviço Cloud Eye da Huawei Cloud

Métricas mensuráveis

Levando em consideração que as atribuições relacionadas ao processo de monitoração de infraestrutura são realizadas de forma automática em ambientes de nuvem, vamos entender como essa dinâmica funciona, para que possamos tirar melhor proveito dos recursos que o paradigma nos disponibiliza.

O serviço de virtualização *ECS*, pertencente ao portfólio de ferramentas oferecidas pela *Huawei Cloud*, disponibiliza, por padrão e de forma contínua, alguns dados ao serviço de monitoração *Cloud Eye* — o que não ocorre com outro serviço de computação muito popular na plataforma, o *BMS*. Essas informações são denominadas "métricas básicas do ECS".

De forma complementar, para que métricas mais específicas relacionadas ao sistema operacional convidado possam ser coletadas, é necessário que um *software* cliente seja instalado em cada instância de virtualização que se deseja monitorar. A identificação atribuída a esse componente recebe a nomenclatura de "agente".

O agente é responsável por coletar uma série de métricas no sistema em que está instalado e repassar essas informações ao serviço de monitoração a que responde, no nosso caso, o *Cloud Eye*. Vale lembrar que os agentes são compatíveis com serviços de computação baseados em virtualização e com arquiteturas *bare metal*.

Na grande maioria dos casos, o *software* em questão já está presente nas principais imagens disponibilizadas e homologadas pelo provedor de nuvem pública de sua escolha, podendo, ainda, ser incluído em novas imagens criadas ou customizadas pelo cliente, de acordo com a documentação oficial disponibilizada pelo *vendor* escolhido.

De forma geral, cada métrica disponibilizada pelas instâncias ou servidores físicos constitui um *dashboard* informativo que permite que cada uma delas seja monitorada com maior atenção, em função

dos riscos ou da estabilidade que, porventura, venha a representar ao atingir diferentes níveis de criticidade.

Sendo assim, é fundamental que o responsável pela gestão da infraestrutura defina não somente as métricas que deseja inspecionar de forma mais detalhada, como também o intervalo de valores considerados críticos no processo de monitoração de um ambiente em produção.

Alarmes e notificações

Uma vez que o *sysadmin* já tem as informações provenientes do processo de monitoração de sua infraestrutura, é necessário passar para a próxima etapa, que consiste na configuração de alarmes.

Os alarmes podem ser definidos como gatilhos, disparados mediante a ocorrência de determinado evento ou condição que se mantém vigente por um período de tempo mínimo preestabelecido pelo administrador.

O acompanhamento contínuo do consumo de recursos de *CPU*, memória *RAM*, espaço em disco e tráfego de rede são exemplos de casos de uso comuns, utilizados com maior frequência pelos usuários.

Para ilustrar algumas aplicações cotidianas, no caso de um volume de armazenamento poderíamos ter uma regra que entraria em vigor no momento em que o dispositivo tivesse 50% de sua capacidade utilizada, notificando ao administrador a necessidade de substituição ou expansão da unidade atual.

Por outro lado, em um hipotético caso de uso envolvendo a monitoração de recursos de *CPU* de uma instância ou servidor *bare metal*, temos um cenário um pouco mais elaborado. Além da métrica de carga utilizada como gatilho — por exemplo, 60% da capacidade de processamento da *CPU* — é necessário que ela mantenha-se vigente por um intervalo de tempo de 5 minutos, por padrão, que também está especificado na regra do alarme.

Esta particularidade evita que eventuais contramedidas sejam aplicadas de forma imediata, evitando ações potencialmente equivocadas ou precipitadas. Um pico isolado de demanda computacional, proveniente de acessos simultâneos a um serviço hipotético, poderia manter a taxa de processamento acima da média apenas por alguns segundos, o que caracterizaria um alarme falso positivo.

Maiores detalhes sobre este caso de uso em particular serão abordados no capítulo 10, que descreve o recurso de *Auto Scaling*, responsável por promover a implementação da escalabilidade horizontal de infraestrutura de forma dinâmica, ou simplesmente, elasticidade.

Dessa maneira, se bem utilizados, os alarmes podem caracterizar um mecanismo efetivo de resposta a incidentes e tolerância a falhas, permitindo que as respectivas contramedidas sejam executadas em caráter preventivo, mitigando a ocorrência de situações críticas previsíveis.

8.2 Monitoração de serviços

À medida que a infraestrutura de servidores provisionada pelo cliente aumenta, a gestão das informações provenientes do monitoramento dos serviços oferecidos acompanha essa tendência, tornando-se uma tarefa mais trabalhosa, com resultados menos efetivos em curto prazo.

Tal complexidade ocorre não somente em função do número de serviços oferecidos, mas também considerando a existência de eventuais servidores redundantes com vistas ao aumento da disponibilidade e à tolerância a falhas nas instalações da organização.

No entanto, o maior "vilão" nesse tipo de cenário é a natureza distribuída dos registros de *log*, que, por padrão, são gerados e consequentemente analisados de forma isolada, cada qual em um servidor dedicado responsável pela execução do serviço que se deseja monitorar.

Em ambiente de nuvem, é possível lançar mão de serviços específicos e especializados para essa finalidade, com muitas vantagens. Além de permitir a coleta dos *logs* de serviço, essas ferramentas organizam as informações de forma centralizada e disponibilizam mecanismos de consulta baseados em linguagem *SQL* que retornam rapidamente os resultados da pesquisa, independentemente do tamanho da base de dados existente.

Essa abordagem permite um acompanhamento muito mais amplo. Ela parte da geração dos *logs* de todos os serviços oferecidos na infraestrutura, em nível de sistema operacional, passa pela ingestão das informações em um armazenamento centralizado e, finalmente, disponibiliza um *dashboard* que permite a análise e visualização dos arquivos existentes. Além disso, possibilita a busca por padrões por meio do uso de expressões regulares, palavras-chaves ou *strings* mais complexas.

Da mesma maneira, os recursos citados podem ser utilizados para a criação de alarmes baseados em palavras-chaves e afins, com o objetivo de identificar eventuais incidentes em tempo real, o que traz contribuições extremamente significativas para esse tipo de contexto.

Na *Huawei Cloud*, a monitoração centralizada dos *logs* de serviços de rede é realizada pelo serviço chamado *"Log Tank Service (LTS)"*.

Figura 8.2: Monitoração de serviços de rede realizada pelo serviço Log Tank Service (LTS) da Huawei Cloud

8.3 Monitoração de usuários

Um dos maiores desafios encontrados po organizações que utilizam produtos e serviços baseados em soluções digitais é a garantia do não repúdio. O termo, também conhecido como "irretratabilidade", assegura que a autoria de execução de uma determinada ação, atribuída a uma entidade física (usuário) ou virtual (aplicação), não possa ser contestada.

Em um ambiente de nuvem pública, as principais interfaces de interação entre os usuários e a infraestrutura provisionada, assim

como com o portfólio de produtos e serviços, baseiam-se em um modelo cliente-servidor.

Neste paradigma, as solicitações são realizadas pelos usuários via console de gerenciamento *web* ou, de forma programática, por meio da utilização das *APIs* de serviço disponibilizadas pelo provedor, consumidas por *scripts* de automação ou aplicações desenvolvidas pelo cliente.

Dessa forma, um eventual incidente ocasionado por ações indevidas de um usuário específico pode ser investigado com o auxílio de ferramentas de monitoração de atividades. Tais recursos, disponibilizados como serviço pelo provedor de nuvem pública, tornam possível a identificação do responsável com precisão, evitando o não repúdio.

Tudo isso é possível graças ao modelo usado nas operações do provedor de nuvem. Como os serviços oferecidos pela plataforma se caracterizam como um conjunto de *web services*, o *software* cliente adotado pelo usuário envia uma solicitação ao serviço que deseja utilizar, que, por sua vez, devolve uma resposta como retorno ao comando enviado inicialmente.

Como o usuário em questão está devidamente autenticado para que possa ter a autorização concedida por meio de políticas de segurança específicas para realizar qualquer tipo de operação, todas as suas ações são automaticamente registradas, com as *timestamps* correspondentes e APIs utilizadas, disponibilizando registros plenamente auditáveis.

Na *Huawei Cloud*, a monitoração das operações realizadas pelo usuário é realizada pelo serviço denominado *"Cloud Trace Service (CTS)"*.

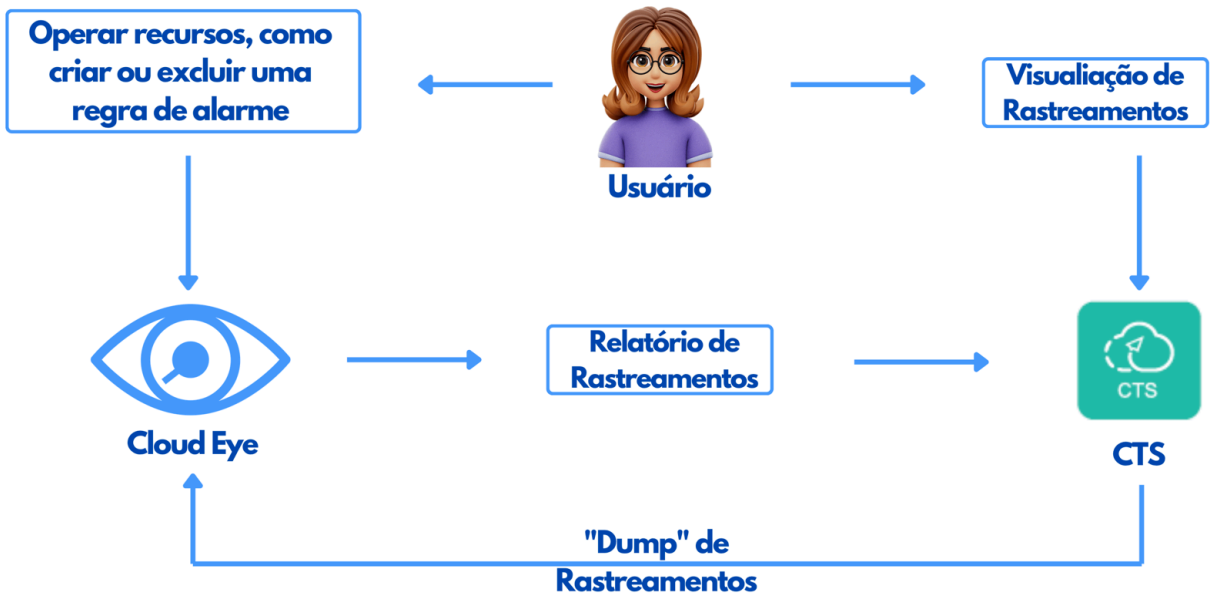


Figura 8.3: Monitoração das operações realizadas pelo usuário via Cloud Trace Service (CTS) da Huawei Cloud

8.4 Considerações importantes

A utilização de serviços de monitoração em ambiente de nuvem é fundamental para que aplicações *cloud native* tirem o melhor proveito possível do paradigma, justificando a adoção do modelo de nuvem e tornando-o competitivo e diferenciado, tanto do ponto de vista tecnológico quanto econômico. Isso beneficia profissionais de desenvolvimento, operações e manutenção e, principalmente, os usuários finais, que terão à sua disposição uma experiência de uso diferenciada.

Provedor	Monitoramento de infraestrutura	Monitoramento de usuários
Alibaba Cloud	Cloud Monitor	Action Trail
Amazon Web Services	Cloud Watch	Cloud Trail
Google Cloud Platform	Cloud Monitoring	Registros de Auditoria do Google Cloud
Huawei Cloud	Cloud Eye	Cloud Trace Service
IBM Cloud	IBM Instana Observability	IBM Cloud Logs
Microsoft Azure	Azure Monitor	Azure Monitor
Oracle Cloud Infrastructure	OCI Monitoring	-

Questionário online

Figura 8.4: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 8.5: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo

CAPÍTULO 9
Serviços gerenciados pelo provedor de nuvem

Ao tomar a decisão de provisionar infraestrutura em ambiente de nuvem, o cliente deve ter em mente que existem diferentes modelos de serviço disponibilizados por esse paradigma, cada um com suas particularidades, conforme discutido no capítulo 2.

- Infraestrutura como Serviço *(IaaS)*;
- Plataforma como Serviço *(PaaS)*;
- *Software* como Serviço *(SaaS)*.

O principal diferencial entre as modalidades mencionadas refere-se, sobretudo, aos impactos que cada uma das categorias gera no chamado "modelo de responsabilidade compartilhada".

Naturalmente, em ambientes em que o *datacenter* encontra-se em infraestrutura própria, o cliente é responsável por todas as operações realizadas, independentemente da camada em que são executadas. Tal cenário exige que as organizações tenham equipes de colaboradores fixas, numerosas e altamente especializadas — fatores que impactam significativamente suas despesas operacionais mensais e, consequentemente, o custo total de propriedade.

Considerando a diversidade de profissionais necessária nos processos desenvolvidos nos setores de infraestrutura, análise, desenvolvimento, segurança e banco de dados — para citar apenas alguns exemplos — é preciso atenção redobrada por parte dos responsáveis por um eventual processo de migração para ambiente de nuvem, caso a principal motivação da transição de paradigma esteja concentrada na expectativa específica de redução de custos com a folha de pagamento da organização.

Figura 9.1: Modelo de Responsabilidade Compartilhada

Em um modelo de contratação *IaaS*, no qual apenas a gestão da infraestrutura de *hardware* e virtualização é de responsabilidade do

provedor de nuvem pública, apenas as equipes desses dois setores estariam aptas a passar por um eventual processo de redução.

Todos os demais profissionais continuam necessários e mantêm a mesma carga de trabalho e responsabilidades do modelo baseado em infraestrutura própria, considerando que o exemplo em tela retrata uma estratégia de migração *lift and shift*, ou seja, um modelo que não requer alterações significativas nos fluxos de trabalho vigentes adotados até então.

Em um primeiro momento, essa opção de *rehosting* pode ser interessante para tirar proveito de alguns dos benefícios do paradigma da computação em nuvem, especialmente os relacionados à agilidade de implementação e à escalabilidade ao alcance de um clique.

No entanto, o modelo em questão pode não ser o ideal para *startups* ou empresas de pequeno porte que não contam com equipes tradicionais de tecnologia da informação, compostas por diferentes profissionais de diferentes áreas e especialidades.

Tal característica, apesar de contornável, poderia inviabilizar economicamente o negócio, a depender da natureza do projeto a ser desenvolvido. Nesse sentido, existem alternativas que permitem ao cliente delegar boa parte de suas responsabilidades ao provedor de nuvem pública, tornando o processo de migração ou implantação ainda mais ágil e factível no curto prazo.

Nos chamados "serviços totalmente gerenciados pelo provedor", como o próprio nome sugere, a maior parte das atribuições é destinada ao provedor de serviços de nuvem pública. Ao cliente, cabem apenas as responsabilidades localizadas nas camadas superiores da pilha de infraestrutura — geralmente relacionadas aos dados manipulados e à aplicação diretamente utilizada.

9.1 Caso de uso: Implantação de um serviço de gerenciamento de banco de dados relacional

Para exemplificar de maneira mais clara o impacto da adoção de serviços totalmente gerenciados pelo provedor de nuvem pública, vamos recorrer a um caso de uso baseado na utilização de servidores de bancos de dados relacionais. Em uma implementação hipotética, realizada inicialmente a título de prova de conceito ou utilizando uma estratégia de migração *lift and shift*, o cliente certamente poderia lançar mão de uma instância provisionada pelo serviço de virtualização *Elastic Cloud Server (ECS)*, usando o modelo de contratação sob demanda.

O serviço em questão deve, portanto, implementar um sistema de gerenciamento de banco de dados em um serviço de virtualização com configurações de *hardware* compatíveis com as necessidades definidas inicialmente pelo solicitante do projeto. Tendo em vista as características do serviço de computação escolhido, podemos considerar que o sistema operacional adotado poderia ser baseado em uma distribuição *Linux* da preferência do cliente, disponibilizada pelo provedor de nuvem pública por meio de uma imagem homologada pelo serviço *Image Management Service (IMS)*, por exemplo. No cenário proposto, o cliente seria responsável por toda a configuração do sistema operacional selecionado, bem como do sistema de gerenciamento de banco de dados escolhido e dos demais serviços correlatos. Todas as configurações relacionadas a usuários, grupos, políticas de acesso, segurança e *patches* de correção também devem ser administradas por profissionais pertencentes à organização.

Não podemos esquecer de atribuições de extrema relevância, como aquelas relacionadas à monitoração de *logs*, configuração de disponibilidade, tolerância a falhas, rotinas de *backup*, restauração em caso de desastres, escalabilidade e balanceamento de carga,

pois são serviços que requerem ações imediatas em casos de incidente, de modo a preservar a continuidade e disponibilidade dos negócios.

A utilização de serviços como o *Auto Scaling* pode auxiliar o gestor nas questões envolvendo disponibilidade, tolerância a falhas e balanceamento de carga, mas, ainda assim, exigiria pessoal especializado para os processos de configuração e manutenção cotidiana.

Os demais serviços supracitados poderiam ser automatizados via utilização de *scripts* ou serviços específicos disponibilizados pelo provedor de serviço, mas, da mesma maneira, demandariam mão de obra específica.

Nesse sentido, fica claro que, ao adotar o modelo de infraestrutura como serviço, as organizações precisam ter em mente que, além de manter os profissionais de tecnologia em seu quadro de colaboradores, será necessário qualificá-los para que possam atuar de forma eficiente no novo paradigma, executando as tarefas que anteriormente realizavam em um ambiente de infraestrutura própria.

Utilizando o Relational Database Service (RDS)

A seguir, analisaremos uma alternativa de implantação bastante utilizada em ambientes de nuvem, que substitui a metodologia tradicional apresentada na seção anterior.

De acordo com a *Huawei Cloud*, o *Relational Database Service (RDS)* é um *web service* baseado em nuvem considerado confiável, escalável, performático, seguro, simples de gerenciar e imediatamente pronto para uso. Além disso, alia todas essas características a um ótimo custo-benefício.

O *RDS* suporta os principais *engines* de sistemas de gerenciamento de bancos de dados utilizados no mercado, a saber:

- MariaDB;

- MySQL;
- PostgreSQL;
- SQL Server.

Dessa forma, ao optar por uma das diferentes soluções de gerenciamento de banco de dados disponibilizadas pelo serviço, o cliente pode eventualmente dar preferência a tecnologias com as quais já possua alguma afinidade ou que, porventura, já tenham sido utilizadas em sua infraestrutura *on-premise*.

Após a definição da solução tecnológica, os profissionais responsáveis por tarefas especializadas — especialmente administrativas e de manutenção pertinentes à gestão de infraestrutura —, bem como pelos processos de *performance tuning* do sistema de gerenciamento de banco de dados adotado, poderão desempenhar suas atividades por meio do console de gerenciamento do serviço em questão, neste caso, o *RDS*.

Da mesma maneira, desenvolvedores e usuários finais da aplicação disponibilizada em ambiente de nuvem poderão executar suas atribuições interagindo diretamente com a aplicação na camada de acesso, cada qual com seus papéis específicos.

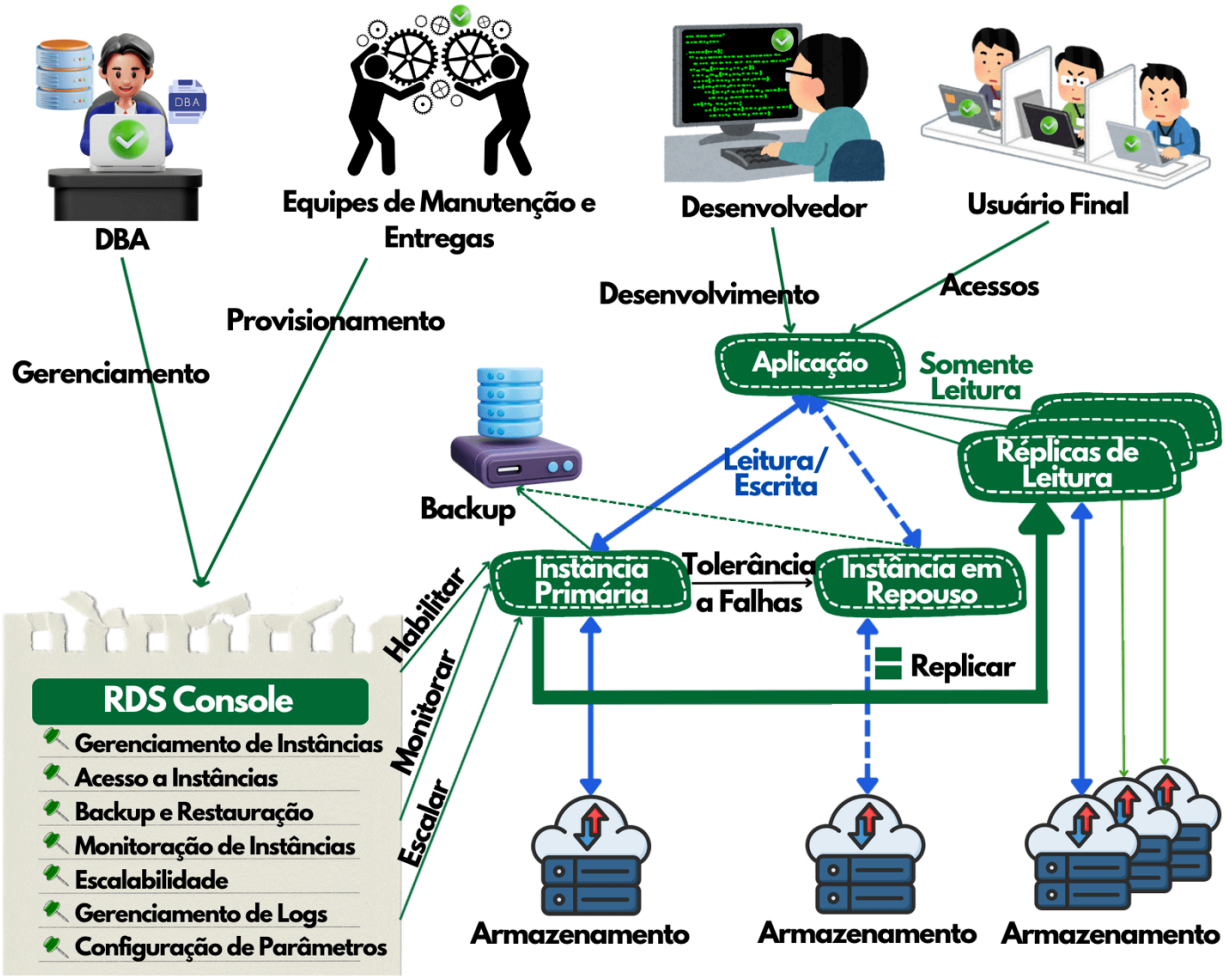


Figura 9.2: Conexão de instância de banco de dados

Vale lembrar que *engines* de bancos de dados *open source* suportados pelas soluções totalmente gerenciadas ofertadas pelos provedores de nuvem pública normalmente recebem otimizações que as tornam mais performáticas do que uma solução padrão, sem modificações.

Para efeito de comparação, em cenários de alta concorrência, o produto *RDS for MySQL* da *Huawei Cloud* apresenta desempenho três vezes superior em relação a uma solução equivalente *on-premise* ou executada na modalidade de infraestrutura como serviço, tendo em vista que a solução em questão utiliza um *kernel MySQL* aprimorado da *Huawei*, denominado *HWSQL*.

Adicionalmente, as instâncias provisionadas pelo serviço são isoladas em sub-redes privadas, com configurações específicas de grupos de segurança, de modo a restringir o acesso aos dados somente a origens conhecidas e confiáveis.

Backup, restore e ferramentas de migração de dados

O *RDS* suporta *backups* automáticos, manuais e períodos de retenção de dados customizados. A opção de automação da rotina de *backups* é positiva, pois elimina a dependência da interação humana em um processo cotidiano e repetitivo, que eventualmente pode falhar ou ser colocado em segundo plano esporadicamente.

Nesta modalidade, os *backups* são gerados diariamente e armazenados no *Object Storage Service (OBS)* por um período de 1 até 732 dias, o que representa um período de retenção de aproximadamente dois anos. Este período, definido por padrão para 7 dias na política de *backups* automáticos, pode ser customizado pelo cliente e está diretamente associado ao tempo de armazenamento dos arquivos em questão, tendo em vista que não podem ser removidos manualmente.

É possível ao usuário criar políticas de *backups* personalizadas de acordo com sua necessidade específica, ou ainda realizar *backups* isolados sob demanda, de forma manual.

É importante que o cliente tenha em mente que, a depender do tamanho de sua base de dados, o custo de armazenamento dos arquivos no sistema de armazenamento de objetos por um período de tempo tão grande pode representar uma despesa mensal significativa. Nesse sentido, vale lembrar que existem estratégias que podem ser utilizadas para reduzir este custo, como a personalização de classes de armazenamento pertinentes aos *backups* durante o seu ciclo de vida, por exemplo.

É possível restaurar *backups* facilmente, com apenas alguns cliques, em qualquer momento com data e hora compreendidos no período

de retenção vigente. As janelas de recuperação têm intervalos mínimos de 5 minutos, referentes à periodicidade em que os *backups* incrementais são realizados.

O processo de restauração pode contemplar instâncias completas ou tabelas de bancos de dados específicas, com destino a novas instâncias *RDS* ou mesmo à instância original.

A maioria dos provedores de nuvem pública oferece soluções para migrar dados de um sistema de gerenciamento de banco de dados funcional, geralmente provisionado na infraestrutura *on-premise* do cliente, para soluções gerenciadas na nuvem. Da mesma maneira, por utilizarem *engines* amplamente adotados no mercado, esses serviços também são compatíveis com as ferramentas de *backup* e *restore* nativas ou comerciais normalmente utilizadas fora do ambiente de nuvem.

A principal solução de migração de dados oferecida pela *Huawei Cloud* é a *Data Replication Service (DRS)*, ferramenta capaz de realizar a tarefa de migração de toda uma base de dados em questão de minutos, de modo simples e intuitivo.

Recursos adicionais

Diferentemente das soluções *on-premises* ou das alternativas disponíveis no modelo de infraestrutura como serviço, nas soluções totalmente gerenciadas ofertadas pelos provedores de serviço, recursos de segurança, disponibilidade e balanceamento de carga também são contemplados.

Inicialmente vamos tratar sobre o provimento de recursos de alta disponibilidade para o serviço de banco de dados. Para que a funcionalidade em questão esteja disponível, é necessária a criação de uma cópia idêntica à da instância *RDS* de origem, denominada como "primária". A réplica recém-criada recebe o nome de *"standby"* e, em geral, é preferencialmente provisionada em uma zona de

disponibilidade diferente daquela em que a instância original está localizada.

Para manter a réplica atualizada, a instância primária envia atualizações síncronas para a instância de destino automaticamente, toda vez que uma instrução de gravação ou atualização é realizada.

O cenário em questão implementa uma solução nativa de tolerância a falhas. Para um melhor entendimento da arquitetura proposta, é importante que se saiba que cada uma das instâncias possui um endereço *IP* privado de gerenciamento próprio, utilizado para manutenções e demais tarefas administrativas, bem como um segundo endereço privado de serviço, comum e compartilhado entre a instância primária e a *standby*, utilizado como *IP* de destino pela aplicação.

Nesse contexto, em um primeiro momento a instância primária responde pelo *IP* de serviço e atende às requisições encaminhadas pela aplicação. Em caso de falha da instância principal, automaticamente a réplica em *standby* assume o endereço de serviço, de forma transparente para o cliente e para a aplicação, tendo em vista que não existe a necessidade de alteração de endereçamento da base de dados.

O recurso de alternância de endereçamento entre as instâncias é possível em função de "checagens de saúde", realizadas de forma periódica pelo serviço, entre as instâncias primária e *standby*, via rede.

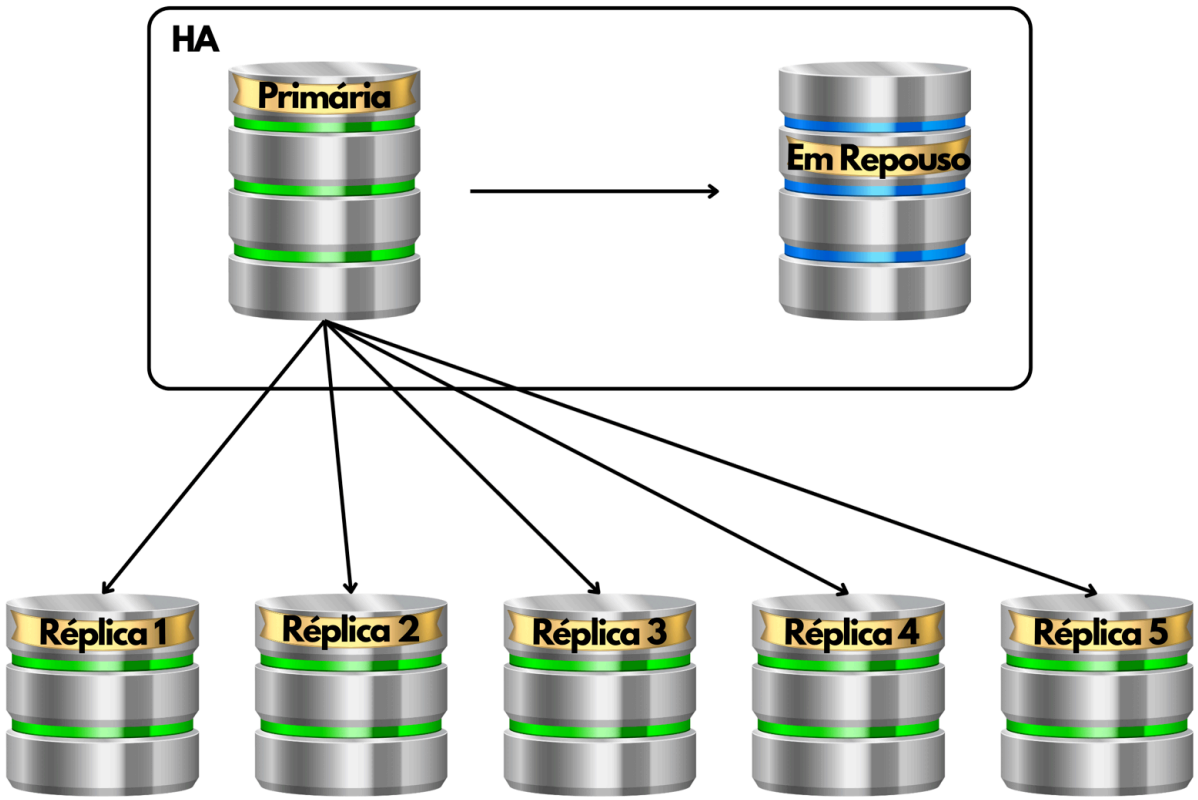


Figura 9.3: Conexão de instância de banco de dados

De forma complementar, é possível lançar mão de outro recurso relevante disponível no *RDS*, capaz de diminuir a carga de trabalho atribuída a uma instância primária, caso necessário.

Em aplicações onde a demanda de consultas ao sistema de gerenciamento de banco de dados é constantemente alta ou apresente tendência de crescimento ao longo do tempo, é importante que a arquitetura utilizada considere uma solução de balanceamento de carga, de modo a evitar gargalos que impactem sua performance e estabilidade.

A recomendação supracitada pode ser implementada facilmente, por meio da criação de até cinco *"read replicas"* adicionais, utilizadas somente para operações de leitura, como o nome sugere. A realização das instruções de escrita permanece sob responsabilidade

exclusiva da instância primária, de modo a facilitar a manutenção da consistência de dados no cenário proposto.

Uma topologia completa do serviço adiciona ainda uma camada extra que atua como *proxy*, encaminhando as solicitações de escrita para o *cluster* de alta disponibilidade e as de leitura para o *cluster* de balanceamento de carga, composto exclusivamente pelas réplicas de leitura.

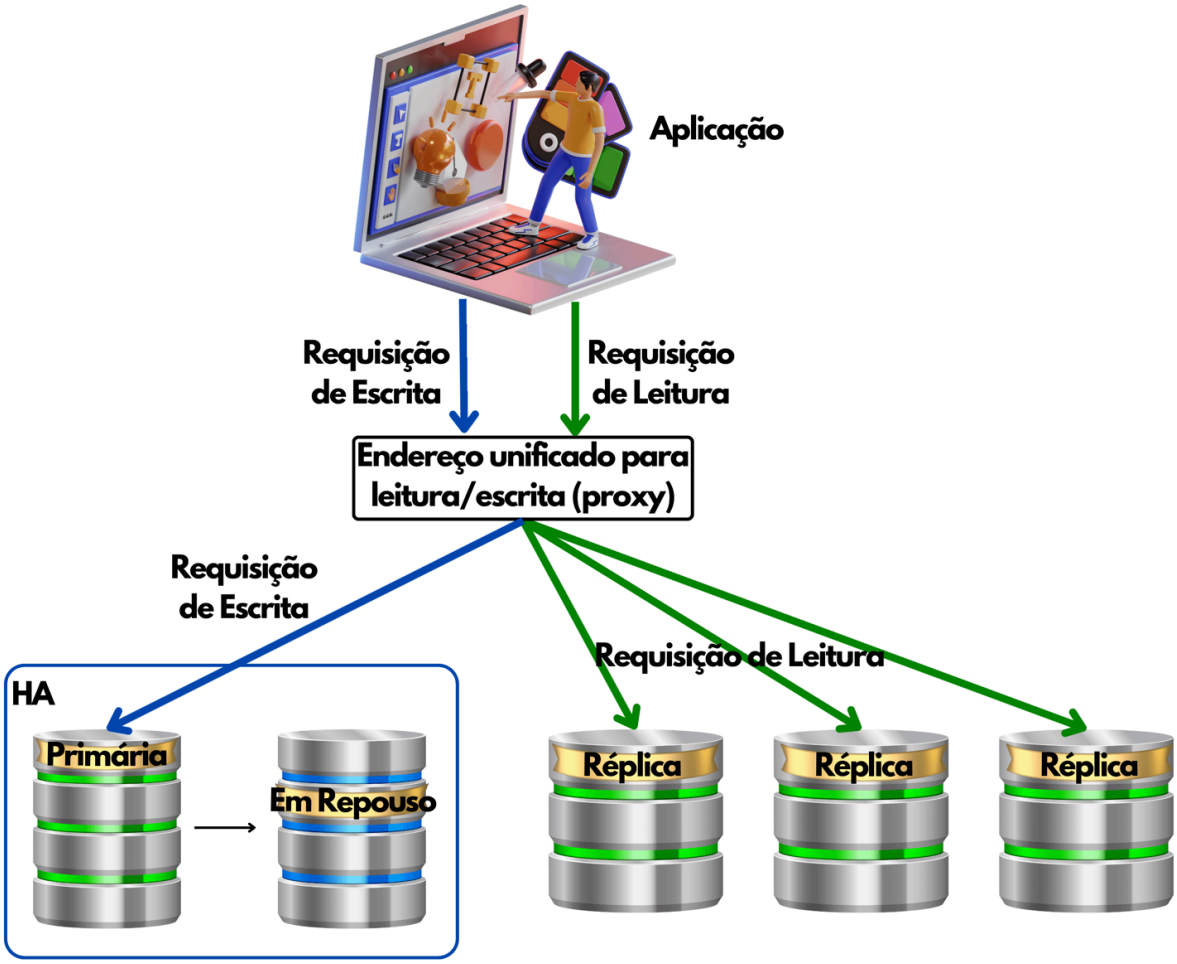


Figura 9.4: Conexão de instância de banco de dados

9.2 Considerações importantes

A utilização de sistemas de gerenciamento de bancos de dados totalmente gerenciados pelo provedor de nuvem pública é apenas um dos exemplos de soluções oferecidas nesta modalidade de serviço.

Levando-se em conta a variedade e, em especial, a não uniformidade de produtos existentes nos portfólios dos provedores de serviço que ocupam esta mesma prateleira, a escolha pela categoria de bancos de dados relacionais apresentada no presente capítulo justifica-se pela existência de oferta de aplicações desta natureza por parte de cada um dos principais provedores de nuvem atuantes no mercado.

Nesse sentido, é importante ter em mente que muito provavelmente teremos ainda mais serviços totalmente gerenciados pelo provedor de nuvem à medida que o tempo passa, tendo em vista o valor agregado das soluções em questão e a agilidade que proporcionam aos clientes.

Provedor	Serviços de bancos de dados relacionais	Serviços de bancos de dados não relacionais
Alibaba Cloud	ApsaraDB RDS, PolarDB	ApsaraDB
Amazon Web Services	Aurora, RDS	DynamoDB
Google Cloud Platform	Cloud SQL	Bigtable
Huawei Cloud	GaussDB, RDS	Document Database Service
IBM Cloud	IBM Cloud Databases	IBM Cloudant
Microsoft Azure	Azure SQL	Azure Cosmos DB
Oracle Cloud Infrastructure	Oracle Autonomous Database	Oracle NoSQL Database

Questionário online

Figura 9.5: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 9.6: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo

Parte 5: Segurança e laboratórios práticos

Capítulo 10
Segurança e conformidade

Os dados de uma organização são o seu maior patrimônio. Garantir a segurança das informações é um pré-requisito fundamental, necessário não somente para atrair novos clientes, mas também para manter a confiança do público-alvo desejado e, consequentemente, a continuidade dos negócios.

Uma das principais dúvidas relacionadas ao paradigma de computação em nuvem refere-se à viabilidade de uso de um modelo cuja localização da infraestrutura de computação e a residência dos dados correspondentes não são conhecidas nem palpáveis. Trata-se de uma questão natural, que na verdade reflete uma boa prática de segurança, a qual impõe restrições severas de acesso aos espaços físicos do provedor de serviços.

É importante termos em mente que não existe uma "bala de prata" definitiva que resolva todo e qualquer problema de segurança. Na verdade, assim como em ambientes *on-premises*, o provisionamento de infraestrutura em ambiente de nuvem exige que as organizações adotem um modelo de segurança baseado em camadas, onde as soluções se complementam para mitigar a ocorrência de incidentes ou, quando eles ocorrerem, reduzir seu alcance e os danos causados. Partindo desse princípio, vale lembrar que existem padrões regulatórios de segurança e proteção de dados amplamente utilizados pela indústria, tais como *HIPAA*, *PCI DSS* e *ISO 27001*. Uma vez adotados e corretamente aplicados, esses padrões permitem à instituição homologar sua infraestrutura e sistemas

conforme as boas práticas de conformidade já validadas por cada uma dessas normas, dentro de suas especialidades.

O modelo de responsabilidade compartilhada adotado pelos provedores de nuvem pública requer atenção redobrada do cliente, especialmente no que diz respeito à conformidade de segurança. O provedor é responsável pela segurança **da nuvem**, ao passo que o cliente é responsável pela segurança **na nuvem**. Em termos práticos, devemos considerar que o cliente utiliza um modelo de contratação de nuvem em que o provedor disponibiliza sua infraestrutura como serviço.

Uma vez que o provedor eventualmente disponha de conformidade com diferentes padrões de segurança amplamente adotados e reconhecidos pelo mercado, vale lembrar que tal característica restringe-se às camadas de responsabilidade do provedor de serviços, neste caso, à camada de *hardware*.

Todas as camadas superiores — incluindo as de gerência, dados e segurança — são de responsabilidade do cliente. Portanto, garantir a conformidade com os padrões desejados acima da camada de *hardware*, de modo a contemplar toda a pilha, também faz parte das atribuições do contratante.

Entender as responsabilidades do cliente e do provedor de nuvem em cada uma das modalidades de serviço disponíveis é de fundamental importância não só para que se possa delegá-las e auditá-las, mas também para garantir a conformidade desejada em toda a pilha.

10.1 A tríade CID

- **Confidencialidade:** a confidencialidade garante que, em um processo de comunicação realizado entre um ponto de origem e

um ponto de destino, somente as partes participantes tenham acesso às informações transmitidas;

- **Integridade:** a integridade das informações transmitidas em um processo de comunicação é satisfeita quando as informações transmitidas na origem não sofreram qualquer tipo de manipulação ou alteração durante o seu trânsito até o destino;
- **Disponibilidade:** a disponibilidade prevê que um recurso computacional ou sistêmico, disponibilizado para acesso público ou privado, esteja acessível e funcional sempre que necessário e a qualquer tempo.

Figura 10.1: Os pilares da segurança da informação

Os serviços e ferramentas de segurança utilizados em ambientes *on-premises* ou de nuvem buscam atender, em maior ou menor medida, a um ou mais dos princípios fundamentais supracitados.

Nas seções a seguir, vamos revisitar alguns tópicos já mencionados em capítulos anteriores e outros que serão apresentados pela primeira vez. A menção aos serviços em questão tem como finalidade dar luz à importância de cada um deles, sem a pretensão de esgotar o assunto ou mesmo de ser uma literatura definitiva sobre o tema.

10.2 Segurança em nível de instância e em nível de sub-rede

O *design* da infraestrutura de um cliente em ambiente de nuvem utiliza tecnologias funcionais que proporcionam o isolamento de recursos privados daqueles que, por definição, devem ser públicos. Uma *Virtual Private Cloud*, ou simplesmente *VPC*, possibilita que um prefixo de endereços *IP* privados seja definido para utilização em um projeto que, por sua vez, distribui esses endereços em diferentes sub-redes, de maneira semelhante à metodologia utilizada nas *LANs* existentes em sistemas *on-premises* tradicionais.

Esta primeira camada de segurança implementa um isolamento lógico entre os ativos do cliente e qualquer elemento do mundo exterior, que pode ser a internet ou mesmo outra *VPC* pertencente ao mesmo cliente ou a um outro qualquer que, porventura, também tenha sua infraestrutura provisionada na mesma região. Para que um recurso computacional provisionado em uma sub-rede possa se comunicar com outros equipamentos fora da *VPC*, são necessários serviços de *gateways* de conectividade, cuja funcionalidade varia de acordo com a necessidade do cliente, conforme vimos no capítulo 7.

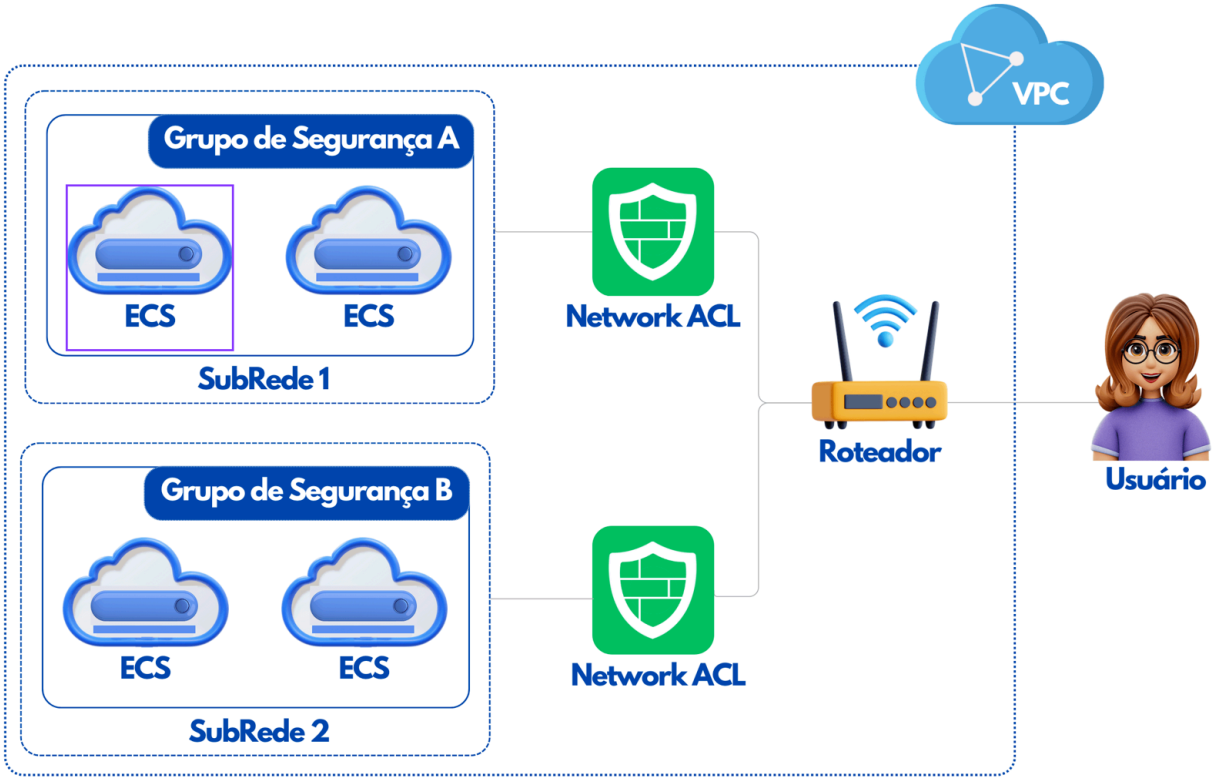


Figura 10.2: Isolamento lógico entre os ativos de nuvem e a internet pública promovido pela utilização da VPC

Para que camadas adicionais de segurança sejam incorporadas à topologia e reforcem o isolamento lógico proporcionado pela *VPC*, o cliente pode ainda lançar mão de recursos de filtragem de pacotes no nível de rede e no nível de cada instância, representados, respectivamente, pelas listas de controle de acesso, ou simplesmente *ACLs*, e pelos grupos de segurança.

Os grupos de segurança aplicam o princípio do privilégio mínimo ao analisar o tráfego entrante em uma instância. Em termos práticos, trata-se de uma política de acesso fechada, onde nenhum pacote vindo de redes externas é aceito. Os grupos de segurança atuam, portanto, como um filtro que analisa regras explícitas que descrevem as condições exatas que vão permitir, em caráter de exceção, que pacotes com determinadas características sejam aceitos.

Em geral, os parâmetros utilizados para descrever tais condições envolvem endereços de origem, bem como protocolo, porta e endereço de destino. A utilização de um grupo de segurança afeta somente a instância na qual está aplicado, sem interferir em instâncias vizinhas.

As *ACLs* possuem um método de funcionamento semelhante, no entanto, são aplicadas a toda uma sub-rede e não a uma instância específica. Nesse sentido, toda e qualquer instância que esteja provisionada em uma sub-rede que tenha uma *ACL* em vigor sofrerá os efeitos que as restrições impostas pelo recurso implementam.

Como dito anteriormente, os recursos de segurança se somam em diferentes camadas. Dessa forma, é possível conciliar a funcionalidade promovida pelas *ACLs* com a desempenhada pelos grupos de segurança, de forma complementar.

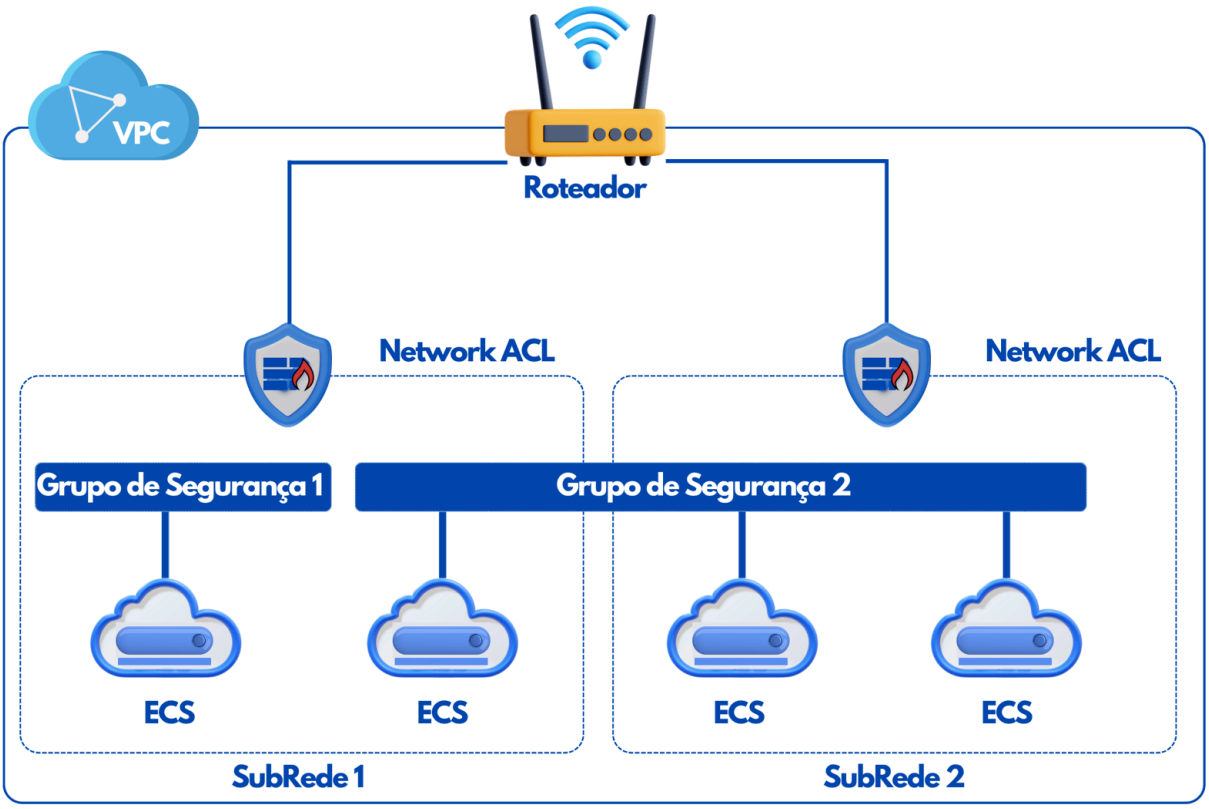


Figura 10.3: Camadas de segurança proporcionadas pela utilização conjunta de VPCs, security groups e ACLs

Tipos de autenticação de instâncias

A administração remota de servidores de rede é algo amplamente utilizado por administradores de sistemas. Ao realizar esse tipo de prática em ambientes controlados ou mesmo por meio da infraestrutura pública da internet via *VPNs*, é possível mitigar vários riscos que se tornam problema em servidores endereçados com *IPs* públicos.

A autenticação padrão de um sistema baseia-se na utilização de um usuário válido e sua respectiva senha de acesso. Ao colocar um servidor *sshd* na internet, ele passa a aceitar conexões de clientes, que podem ser usuários legítimos ou tentativas de acesso não autorizadas, promovidas por ataques automatizados de força bruta

que, em geral, iniciam-se alguns minutos após o servidor ter sido provisionado.

Existem técnicas para mitigar este tipo de ataque. A estratégia adotada pela grande maioria dos provedores de nuvem pública consiste em desativar a autenticação padrão, baseada na utilização de usuário e senha, substituindo-a por uma autenticação baseada em um par de chaves assimétricas, que devem ser definidas no momento em que a instância é criada. O usuário que, porventura, já possua suas próprias chaves e deseje reutilizá-las pode realizar o *upload* de sua chave pública, a ser armazenada na instância desejada; do contrário, ambas serão criadas no processo de configuração da instância.

Apesar de óbvio, vale reforçar que o administrador de sistemas jamais deve reabilitar o acesso baseado na utilização de usuário e senha; do contrário, o sistema passa a disponibilizar novamente um vetor de ataque desnecessário.

Em serviços que precisam manter a utilização do modelo de *login* tradicional, é considerada uma boa prática habilitar também a autenticação em duas etapas ou multifator (*MFA*). Uma vez vencida a primeira etapa, o usuário deve então fornecer um segundo fator de autenticação, que, em geral é baseado em um código recebido por e-mail ou *SMS*, por uma sequência gerada por um sistema de gerenciamento de *tokens* baseado em *software* ou *hardware*, ou ainda por algum tipo de dado biométrico, como a impressão digital, reconhecimento facial ou leitura de retina.

Evidentemente, partimos do princípio de que todo usuário titular da conta de nuvem — também denominado *root* e geralmente identificado por usar as credenciais de e-mail como recurso de autenticação — foi substituído por usuários gerados pelo serviço de gestão de autenticação e autorização, normalmente denominado *IAM*.

10.3 Alta disponibilidade de recursos

A disponibilidade é um dos pilares da segurança da informação e tem como responsabilidade garantir que os recursos de qualquer natureza, como aplicações, acesso seguro e autenticado a dados, servidores e afins, estejam disponíveis a qualquer momento, sempre que demandados ou necessários. Os dispositivos físicos utilizados para provisionamento de infraestrutura são, em sua maioria, equipamentos eletrônicos ou meios físicos de transporte que, evidentemente, podem apresentar problemas. Nesse sentido, é necessário que o administrador de sistemas provisione recursos redundantes, de modo a evitar a existência de pontos únicos de falha que, em algum momento, podem tornar sistemas e informações indisponíveis.

Frente a esse cenário, é preciso ter em mente que um sistema redundante não torna a infraestrutura infalível, mas sim resiliente a falhas. No entanto, é necessário também que os sistemas estejam preparados para acionar eventuais réplicas redundantes quando necessário e, o mais importante, de forma automática, tornando a troca imperceptível ao usuário.

Existem dois tipos principais de estratégias para redundância: as que utilizam réplicas somente quando necessário, ou seja, no caso da ocorrência de incidentes, e as que as utilizam para composição de um *cluster* de balanceamento da carga, responsável por escalar horizontalmente a infraestrutura de forma dinâmica, de acordo com as demandas entrantes, em tempo real.

A primeira estratégia, em geral, é utilizada para replicação síncrona de sistemas de gerenciamento de bancos de dados. Os dados são armazenados em uma réplica que fica em *stand by* e é acionada somente quando a principal apresenta falha. Para que o chaveamento seja automático, a "saúde" do servidor primário é monitorada de forma constante por meio da utilização de comandos *SMTP*, solicitações *HTTP* ou via protocolo *ICMP*, por exemplo.

Quando uma falha ocorre, um endereço *IP* secundário, informado ao servidor *DNS* responsável pela resolução de nomes do serviço, passa a ser utilizado na rota como alternativa de *failover*, até que o principal tenha seu funcionamento normalizado e volte a operar.

Outra solução consiste na utilização de um serviço de *Virtual IP*, que conta com protocolos como o *CARP* ou seus derivados, nos quais cada estação possui seu próprio endereço privado de gerenciamento e ambas compartilham um segundo endereço comum designado ao serviço. Da mesma maneira, no momento em que a máquina principal falha, a segunda, que até então atuava como *backup*, passa a responder pelo *IP* como estação primária, até que o problema seja resolvido, de forma transparente ao usuário final.

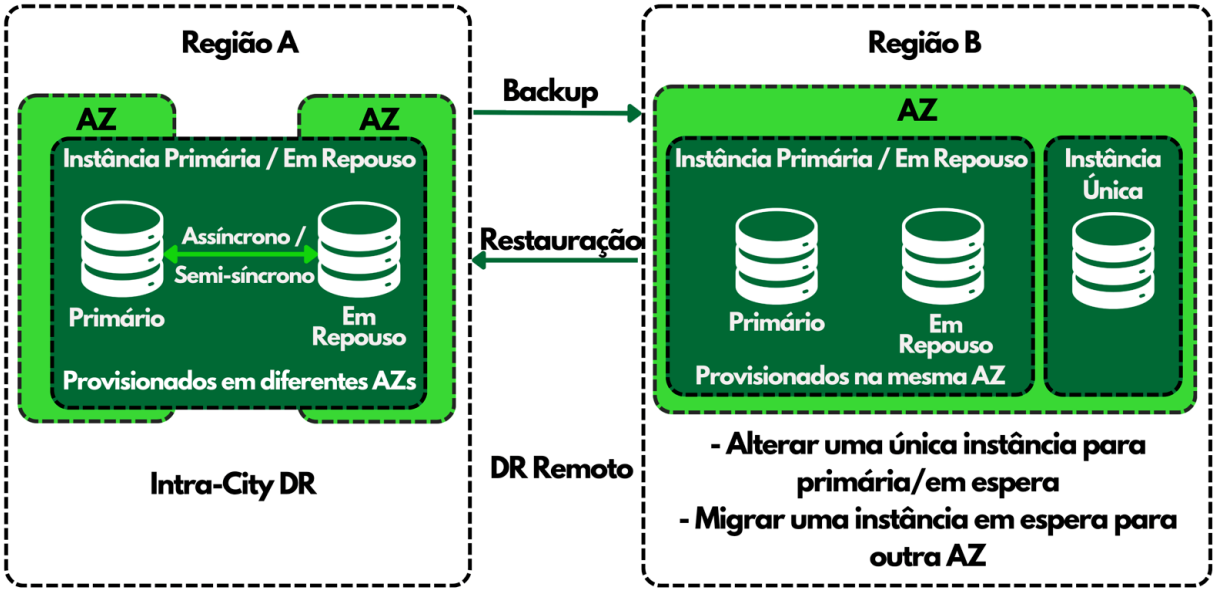


Figura 10.4: Redundância de serviços com sistemas de gerenciamento de bancos de dados RDS

Nos casos em que a infraestrutura redundante é utilizada para balanceamento de carga, é necessário que todos os nós adicionais façam parte de um mesmo grupo de *hosts*, utilizado para compor um *cluster* com essa finalidade. Na topologia, é necessário também um elemento central de acesso que vai distribuir, de forma

alternada, as solicitações entrantes para cada um dos nós pertencentes ao grupo de *hosts*.

A escolha é feita de acordo com o algoritmo utilizado e pode levar em conta o nó menos utilizado, o que possui menor número de entregas ou mesmo uma alternância simples, um a um. Só serão elegíveis como nós de entrega válidos aqueles que forem considerados "saudáveis", para garantir a eficiência do modelo. Para tanto, todos os nós do grupo de *hosts* são monitorados constantemente.

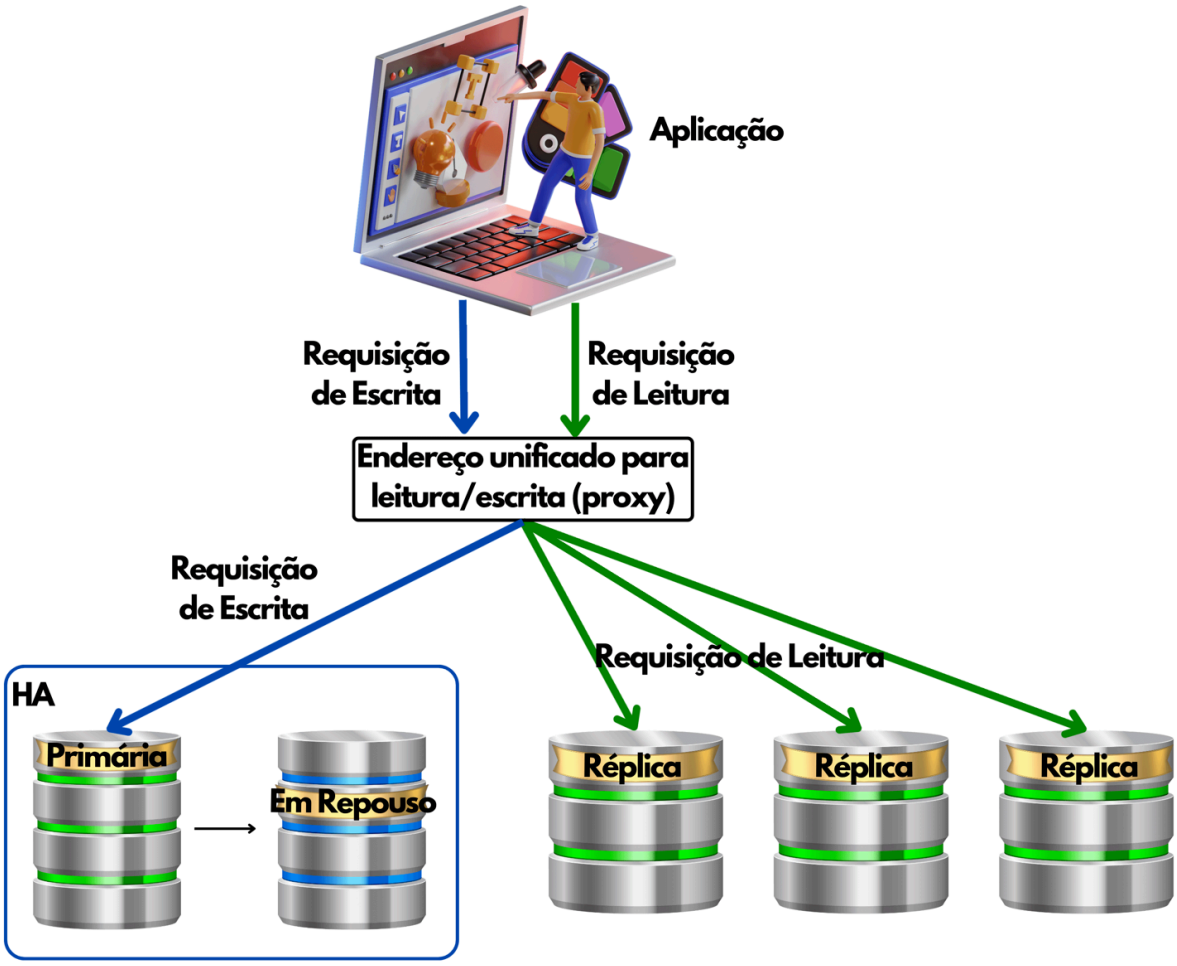


Figura 10.5: Adicionando réplicas de leitura na topologia

Balanceamento de carga, escalabilidade e elasticidade

A escalabilidade de um sistema consiste em sua capacidade de mudar de tamanho, aumentando ou diminuindo sua infraestrutura de acordo com a demanda. Ao optar por provisionar uma aplicação em ambiente de nuvem, é importante que o cliente tenha em mente que alguns conceitos normalmente aplicados em ambientes de infraestrutura própria devem ser revistos.

Uma prática recorrente em soluções *on-premises* tradicionais consiste na cultura da escalabilidade vertical, que, de tempos em tempos, faz com que os componentes principais dos servidores passem por atualizações de configuração, como a troca dos processadores, o aumento da quantidade de memória *RAM* e a utilização de unidades com maior capacidade nos sistemas de armazenamento. Evidentemente, não há nenhum problema em aprimorar a performance dos equipamentos usados. No entanto, ao utilizar este tipo de abordagem é provável que a aplicação fique indisponível em algum momento, seja durante o processo de *upgrade*, que exige o desligamento do equipamento, seja devido a incidentes em arquiteturas sem redundância de servidores, o que caracteriza um ponto único de falha no sistema.

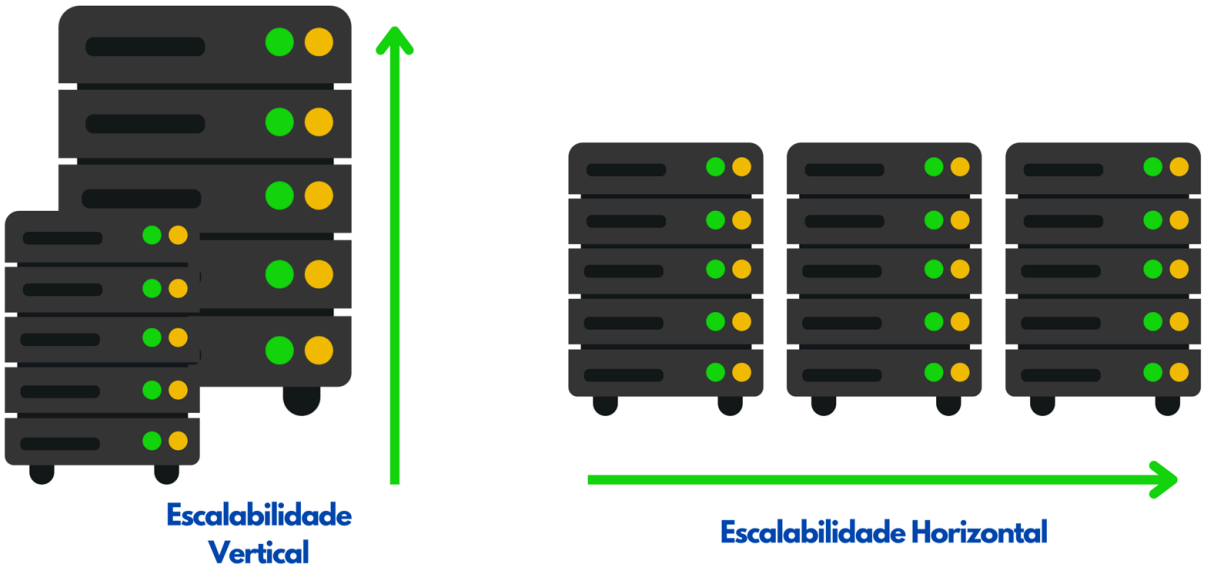


Figura 10.6: Tipos de escalabilidade

Em arquiteturas de nuvem, o modelo preferencial de escalabilidade adotado para garantir a disponibilidade das aplicações é denominado "elasticidade". O conceito consiste na utilização de réplicas idênticas das instâncias de virtualização, organizadas de forma horizontal, cujo ciclo de vida é determinado pela demanda entrante, monitorada automaticamente e em tempo real. Todo o tráfego de acesso é direcionado a um elemento central da arquitetura, denominado "balanceador de carga", um ativo de rede com alta disponibilidade garantida pelo provedor de nuvem. Esse componente é responsável por direcionar cada uma das solicitações, de forma alternada, a um dos nós pertencentes ao grupo de *hosts* que compõem o *cluster* de balanceamento de carga.

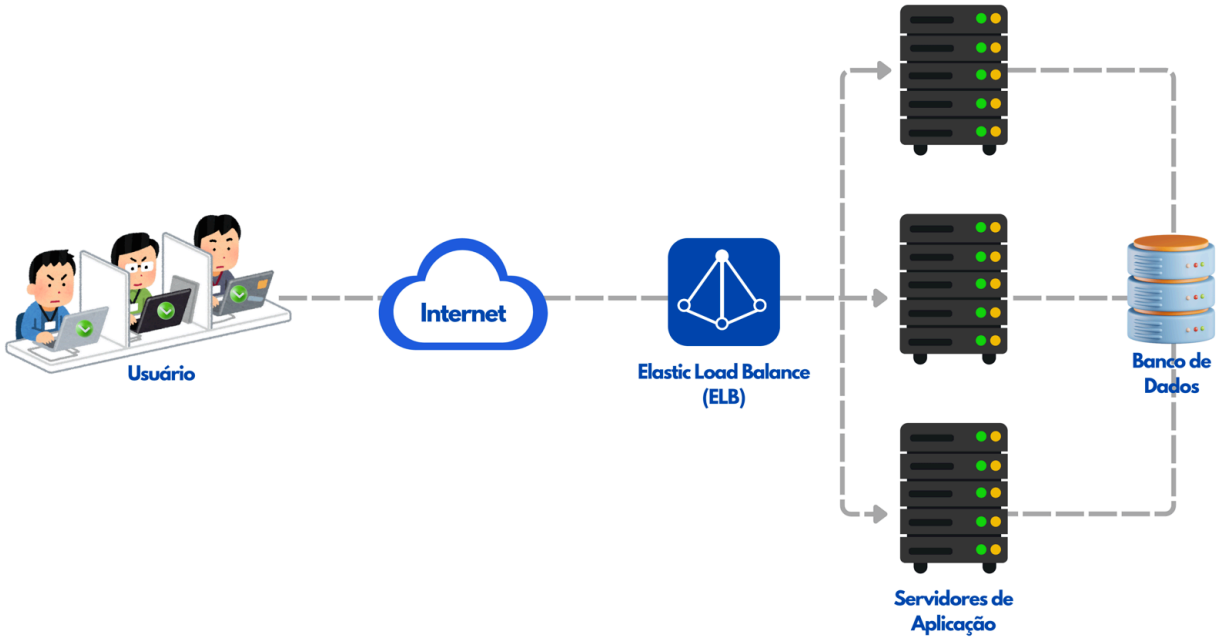


Figura 10.7: Balanceamento de carga do tráfego de entrada

Para garantir que o número de nós em execução no grupo de *hosts* seja sempre o ideal, é necessário que o sistema de monitoramento de infraestrutura entre em ação. Para o caso de uso em questão, a métrica mais indicada consiste na média de utilização de *CPU* de todos os nós pertencentes ao *cluster*. A título de exemplo, se a média de utilização de *CPU* for maior do que 60% durante 5 minutos

consecutivos, aumenta-se um nó no *cluster*, mantendo-se a mesma lógica até que a regra deixe de ser válida. Desse modo, situações atípicas de picos de utilização não esperados deixam de ser um problema.

Da mesma forma, é necessário que a infraestrutura reduza o número de nós, retornando a um patamar suficiente para o atendimento cotidiano assim que os picos de utilização cessarem. Para isso, pode-se configurar uma nova regra baseada na mesma métrica de monitoração de uso de *CPU* mencionada anteriormente: por exemplo, excluir um nó sempre que a média de utilização de *CPU* for menor que 20% durante 5 minutos consecutivos. Vale lembrar que, nesse caso específico, existe uma condição inicial que determina o número mínimo de instâncias que devem permanecer provisionadas, a fim de prevenir que a regra em questão não exclua todos os nós, em caso de um período de ociosidade.

Em geral, o serviço responsável pela implementação do modelo descrito nos provedores de nuvem pública recebe o nome de *Auto Scaling* e atua em conjunto com o sistema de observabilidade oferecido pelo provedor de serviço.

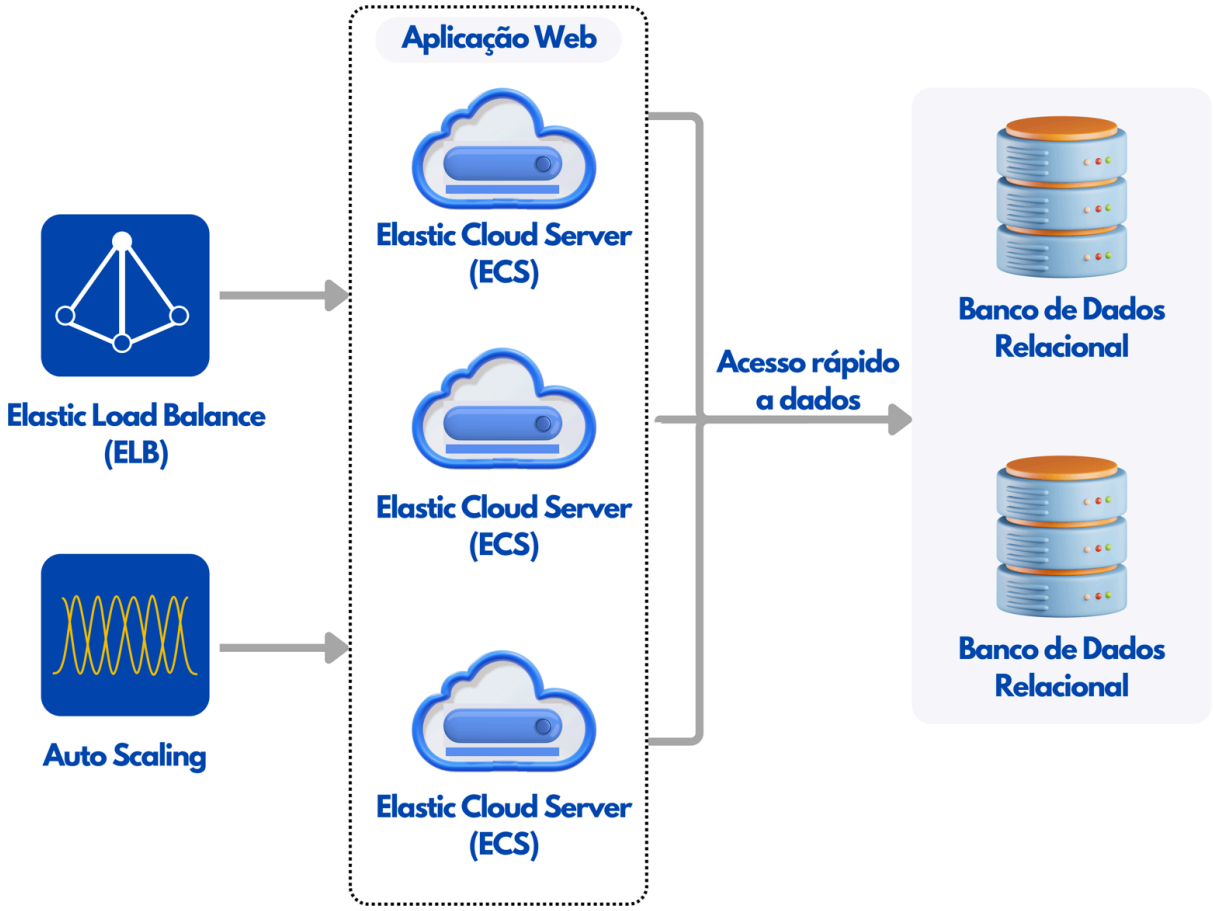


Figura 10.8: Topologia de Auto Scaling aplicada a uma aplicação web típica

10.4 Criptografia e gerência de chaves

Já sabemos que os dados representam os ativos mais importantes de uma organização. Portanto, sua gerência e proteção devem ser uma prioridade constante no cotidiano dos administradores de sistemas.

Existem dois tipos principais de dados: os dados em trânsito e os dados em repouso. Os dados em trânsito são aqueles que estão sendo transmitidos em um processo de comunicação realizado entre um ponto de origem e um ponto de destino, ao passo que os dados

em repouso são aqueles que eventualmente já passaram por essa etapa e estão disponíveis em um sistema de armazenamento no ponto de destino. Em ambas as situações, a utilização de criptografia é altamente indicada, pois, em caso de interceptação das informações por um terceiro não autorizado, os dados estarão cifrados, impedindo o acesso a informações sensíveis.

O uso de criptografia em ambientes controlados pressupõe que se utilizem, preferencialmente, sistemas baseados em chaves assimétricas, compostos por um par de chaves: uma pública e outra privada. Dessa forma, a origem utiliza os dados da chave pública do *host* de destino para encriptar os dados transmitidos, que somente poderão retornar à sua forma original após a efetiva transmissão, por meio da respectiva chave privada.

Para que tudo funcione corretamente, é necessário que o sistema de criptografia utilizado seja forte, tornando inviável a descoberta da chave por tentativa e erro, típica em ataques de força bruta, e garantindo a proteção da chave privada utilizada.

Ao optar pela utilização desse tipo de serviço, o cliente pode administrar suas próprias chaves ou delegar essa função ao provedor de nuvem. Essa escolha, que pode parecer óbvia, pode esbarrar em questões de conformidade ou sistemas legados, levando a organização a optar por gerenciar as chaves utilizadas. Caso isso não ocorra, os provedores de nuvem disponibilizam sistemas de gerenciamento centralizado de chaves que contam, inclusive, com sofisticados recursos especializados baseados em *hardware* para gerar os pares de chaves utilizados, elevando o nível de segurança.

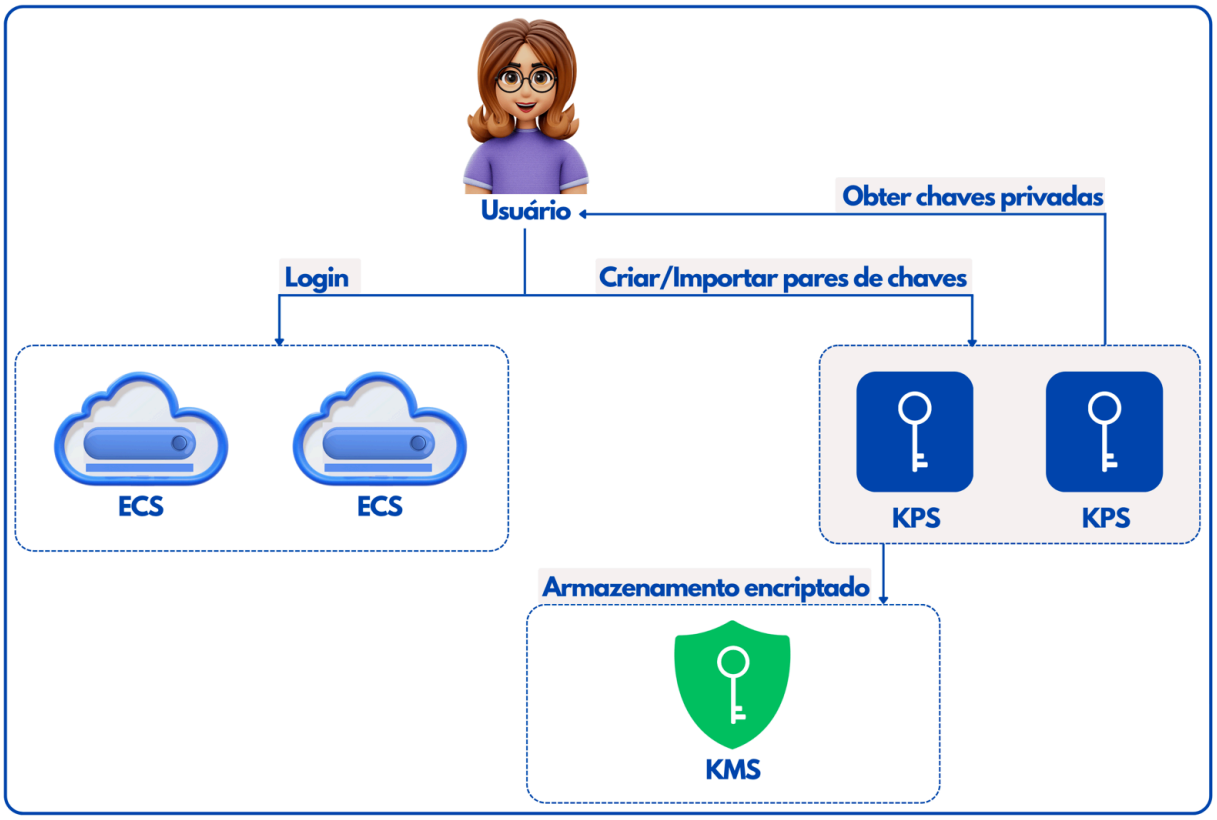


Figura 10.9: Utilização de criptografia e serviços de gerenciamento de chaves assimétricas em ambiente de nuvem

10.5 Appliances de terceiros e o modelo de marketplace

Durante o ciclo de desenvolvimento de uma organização, são várias as soluções de *software* e *hardware* utilizadas até que se encontrem as condições ideais de trabalho e um cenário tido como estável. Para alcançar esse panorama, os colaboradores envolvidos no referido processo certamente destinaram um tempo significativo para que pudessem alcançar o nível de especialização necessário para operar diferentes ferramentas e equipamentos destinados aos sistemas de computação, armazenamento, redes e segurança utilizados internamente.

Nesse sentido, é natural que, em um eventual processo de migração de infraestrutura para ambiente de nuvem, profissionais que até então dispunham de um ambiente funcional e de plena fluência nas soluções que atendiam a contento às suas necessidades e expectativas tenham receio do processo de mudança, pois eventualmente poderão ter que utilizar ferramentas distintas e passar novamente por todo um processo completo de aprendizagem e de avaliação de novas soluções.

A boa notícia é que os principais provedores de nuvem do mercado ofertam a seus clientes várias soluções de *hardware* e *software*, em especial aquelas relacionadas ao gerenciamento de redes e segurança disponibilizadas como *appliances* virtualizadas distribuídas e comercializadas pelo fabricante, por meio do *marketplace* do provedor de nuvem, no formato de imagem. Dessa forma, os colaboradores podem continuar a utilizar os equipamentos de rede gerenciáveis e as soluções de segurança nas quais já possuem notório saber, em sua topologia de nuvem, agilizando a adoção e mantendo a expertise já adquirida anteriormente.

Web Application Firewall

A *Open Web Application Security Project* - *OWASP* (Projeto Aberto de Segurança em Aplicações Web), por meio de seus inúmeros *chapters*, disponibiliza um *ranking* com as dez principais vulnerabilidades da *web* em diferentes partes do globo com periodicidade bianual.

Ao longo dos anos, algumas delas trocaram de posição ou deram lugar a novas falhas. O fato concreto é que, com o passar do tempo, as principais linguagens de programação, sistemas operacionais e sistemas de gerenciamento de banco de dados implementaram mecanismos funcionais para mitigar a ocorrência de tais incidentes. No entanto, a utilização efetiva dos recursos em questão nos

sistemas e servidores ainda segue como uma opção dos desenvolvedores e administradores de sistemas.

Os provedores de nuvem passaram a oferecer soluções capazes de inspecionar o tráfego de pacotes *HTTP* e *HTTPS* com o objetivo de detectar e bloquear os ataques mais comuns, normalmente ranqueados pela *OWASP*, antes que estes possam chegar aos dispositivos finais a que originalmente seriam destinados.

Em geral, tais serviços recebem o nome de *Web Application Firewall*, ou simplesmente *WAF*, nos provedores de serviço.

Host Security Service

Delegar a segurança de um dispositivo final exclusivamente ao bom senso do usuário é uma prática que o administrador de sistemas não pode assumir. Em geral, um típico usuário de tecnologia pode ter o perfil de um simples operador ou mesmo de um hábil desenvolvedor. No entanto, independentemente do perfil técnico, é possível ainda que ele não possua nenhum conhecimento de segurança ou, ainda que tenha, negligencie esse aspecto por nunca ter passado por nenhum incidente durante sua trajetória profissional.

Nesse sentido, é importante poder contar com ferramentas que detectam tráfego malicioso suspeito ou comportamentos anômalos em uma estação de trabalho ou servidor dedicado. Para tanto, basta a instalação de um *software* residente nas estações para que, na ocorrência da detecção de qualquer ameaça, um alarme seja disparado, permitindo que as medidas cabíveis sejam tomadas de forma automática ou semiautomática.

Os agentes, como são chamados, geralmente utilizam um sistema de detecção e/ou prevenção de intrusões que consulta uma base de conhecimento constantemente atualizada que compara binários e tráfego de rede com conteúdos suspeitos a padrões de *malwares* e ataques conhecidos e catalogados localmente ou em rede.

Na *Huawei Cloud*, a gerência dos dados enviados pelos clientes é realizada de forma centralizada pelo *dashboard* do *Host Security Service*, e os agentes precisam ser instalados pelo usuário. Outros provedores de serviço possuem soluções equivalentes e, em geral, já inserem os agentes nas distribuições *Linux* que recomendam prioritariamente.

Anti-DDoS Service

A ocorrência de ataques de negação de serviço distribuídos (*DDoS*) pode comprometer a disponibilidade de uma aplicação, tornando-a inoperante. Um provedor de nuvem disponibiliza várias alternativas para manter a disponibilidade de uma aplicação, especialmente a escalabilidade dinâmica ou elasticidade, baseada na escalabilidade horizontal, que aumenta o número de nós de processamento à medida que a demanda entrante de acessos cresce e balanceia a carga entre os nós.

No entanto, na ocorrência de um ataque desta natureza, o aumento do número de acessos não é proveniente de uma demanda legítima, mas sim de um ataque orquestrado. Dessa forma, aumentar a infraestrutura geraria custos indesejáveis e desnecessários, tendo em vista a natureza do tráfego.

Para mitigar ataques de negação de serviço distribuídos, os provedores de nuvem pública oferecem serviços especializados, denominados *Anti-DDoS*, com suporte a diferentes níveis de proteção que, em geral, podem ser customizados de acordo com as necessidades do cliente.

10.6 Considerações importantes

A segurança da informação é um grande desafio cotidiano para as empresas. O fato de a internet ter se tornado a principal vitrine de

produtos e serviços das organizações fez com que sua oferta naturalmente migrasse para essa nova "prateleira". Esse nível de visibilidade é vital para o sucesso dos negócios, mas perigoso para sua continuidade, caso a exposição seja maior do que o recomendado.

Na prática, estamos falando de um jogo de gato e rato que está longe de acabar. Atualmente, o protagonismo de indivíduos altamente qualificados, responsáveis por ataques sofisticados a alvos predeterminados, convive com o uso de modelos de Inteligência Artificial empregados para a mesma finalidade.

A tendência é que novas ferramentas, desenvolvidas como contramedidas para essas e outras ameaças que surgirem, incorporem ainda mais opções ao portfólio da categoria de serviços de segurança e conformidade oferecidos pelos provedores de nuvem.

É importante que você tenha clareza de que o presente capítulo visa apresentar, de forma superficial, as principais ferramentas de segurança que, inevitavelmente, farão parte do cotidiano de todo profissional que atua em ambientes de nuvem. Existem ainda inúmeros outros serviços especializados voltados ao mesmo fim, que atendem a necessidades e casos de uso específicos. O número de novas soluções tende a aumentar a cada dia, o que exige consulta constante à literatura especializada sobre cada produto, dada a grande quantidade de recursos disponíveis.

Questionário online

Figura 10.10: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 10.11: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo

Capítulo 11
Discussões e conclusões

Ao longo dos últimos dez capítulos, foram apresentados conceitos essenciais relacionados ao paradigma de computação em nuvem. Partindo deste panorama, é possível afirmar que você está apto(a) a operar as ferramentas disponíveis nas principais categorias de serviço em ambiente de nuvem, independentemente do provedor de serviço utilizado.

Evidentemente, ao tratar de temas relacionados a qualquer tipo de tecnologia computacional, é preciso ter em mente que mudanças significativas nas ferramentas utilizadas, em curto ou médio prazo, são inevitáveis, dado o dinamismo da área.

Profissionais da área devem manter-se em um processo de atualização constante, a fim de saber escolher e utilizar, de forma eficiente, as principais soluções e tecnologias empregadas em seu cotidiano profissional. No encerramento de cada capítulo, apresentamos questionários e recursos didáticos que serão revistos constantemente pelo autor. Essa metodologia permite que a contribuição da obra seja ampliada, uma vez que conta com um mecanismo dinâmico de inserção de novos materiais que, na medida do possível, buscam manter o conteúdo sempre atualizado.

Da mesma maneira, nas próximas seções são apresentadas duas plataformas relevantes para que você possa complementar seus estudos. Dentre os conteúdos disponíveis, destacam-se aulas em vídeo, materiais didáticos em formato *PDF*, simulados e laboratórios práticos *online* que dispensam a necessidade de criação de uma conta em um provedor de nuvem pública para a realização de atividades práticas — fundamentais para a fixação dos conteúdos trabalhados.

Ao final deste capítulo, de forma análoga aos demais, temos um questionário que, neste caso específico, tem o objetivo de funcionar como instrumento de avaliação. Ele permite que você compartilhe suas experiências e impressões ao longo da leitura, contribuindo para a melhoria e ampliação dos conteúdos em edições futuras, inclusive com sugestões.

Por fim, na última seção, você terá acesso a recursos de aprendizagem adicionais, disponibilizados pelo programa acadêmico apresentado ao longo do presente capítulo, do qual o autor é instrutor credenciado e certificado.

11.1 Plataforma Huawei Talent Online

A plataforma *Huawei Talent Online* (https://e.huawei.com/en/talent/portal/#) é reconhecida em todo o mundo como um grande repositório de recursos de capacitação e certificação, voltado, em especial, ao desenvolvimento de talentos.

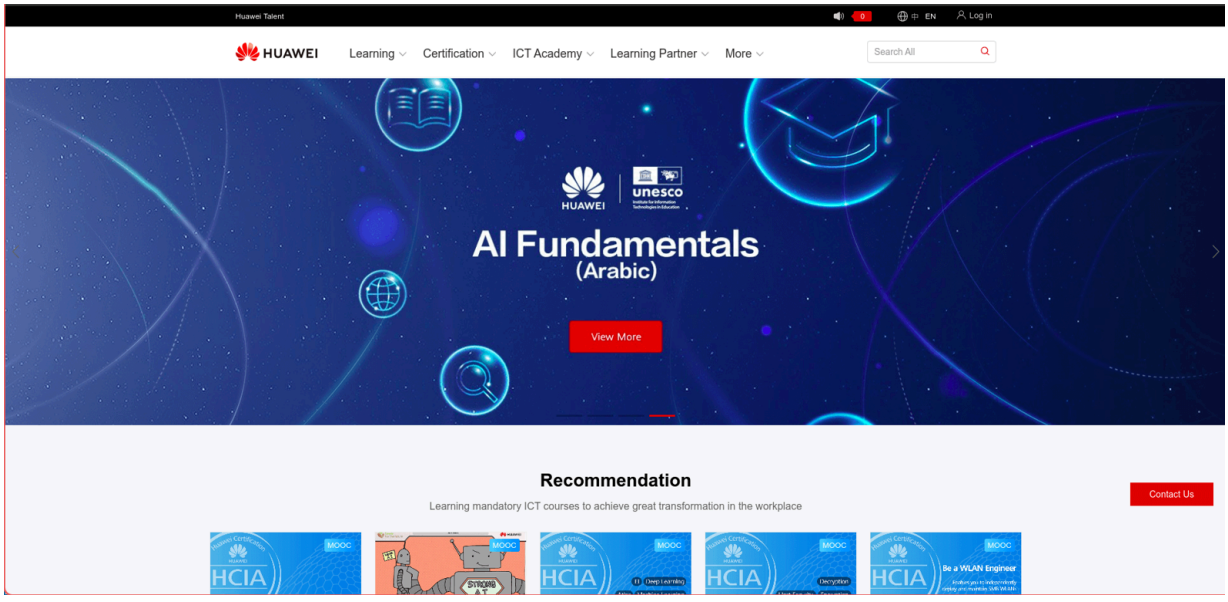


Figura 11.1: A plataforma Huawei Talent

Ao acessar a página principal, o usuário tem a possibilidade de explorar seções específicas disponíveis no menu superior da interface web ou pelo menu representado por três barras horizontais empilhadas, caso esteja utilizando um dispositivo móvel.

A seguir, serão apresentadas algumas das opções encontradas no menu supracitado, de acordo com a aderência proposta por este capítulo:

Learning:

Nesta seção, o visitante pode pesquisar por capacitações que eventualmente tenha o interesse em realizar. Para uma melhor experiência, é recomendado que o usuário utilize os filtros de busca disponíveis, considerando a grande quantidade de treinamentos ofertados na plataforma.

É possível refinar os resultados retornados de acordo com a área, setor da indústria, nível de certificação ou tecnologia a que se referem os treinamentos, bem como por tipo, modalidade, nível de dificuldade, custo e idioma dos treinamentos.

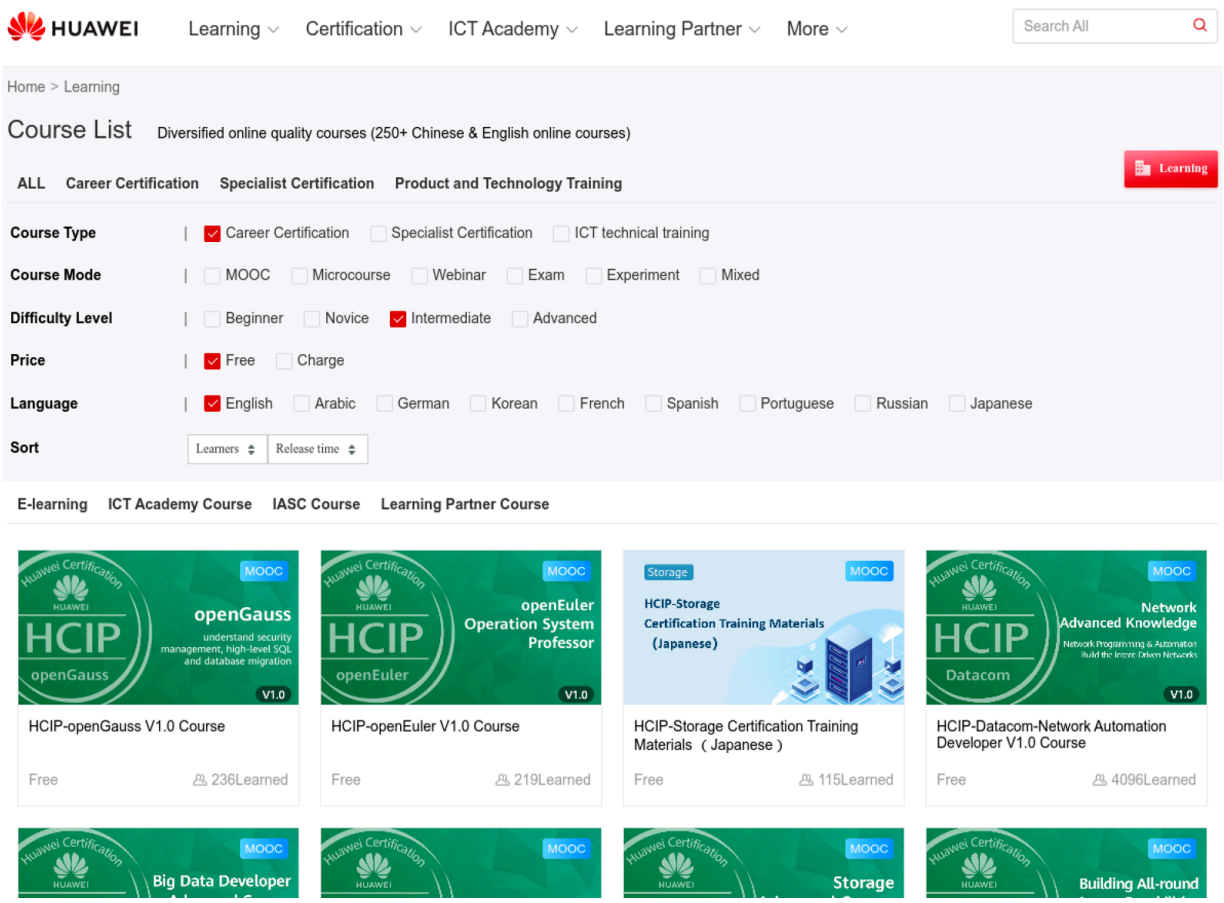


Figura 11.2: Lista de treinamentos disponíveis na plataforma Huawei Talent

Uma vez encontrado o treinamento desejado, já é possível ter acesso a informações preliminares referentes ao conteúdo do curso em questão, ainda que de forma resumida. Para tanto, basta que o candidato ou candidata clique no *banner* ou título de seu interesse. Vencida esta etapa e atendidas as expectativas de ementa, o(a) visitante poderá, então, se inscrever em uma turma vigente e ter acesso imediato aos materiais didáticos e demais recursos disponibilizados.

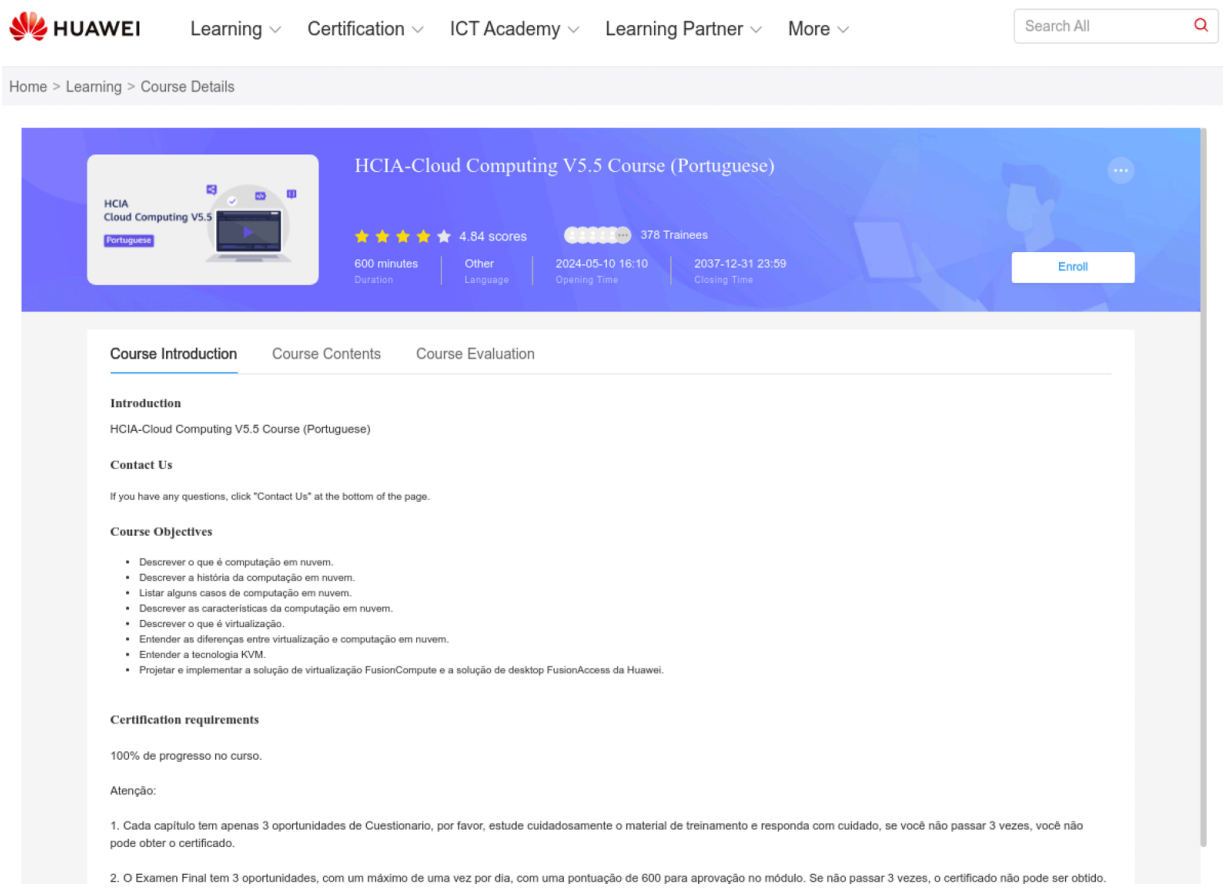


Figura 11.3: Detalhes resumidos de um dos treinamentos ofertados na plataforma

Certification:

Dentre os treinamentos disponibilizados na plataforma *Talent*, uma parcela significativa é destinada à preparação de candidatos e candidatas que desejam se submeter a um exame de certificação profissional oferecido pela *Huawei* e gerenciado por centros de aplicação de testes parceiros. Esta seção possui cinco módulos:

- ***Visão Geral:*** traz informações essenciais de forma objetiva, que auxiliam o candidato a rapidamente compreender a estrutura, os níveis de especialização e demais particularidades do sistema de certificação da *Huawei*;

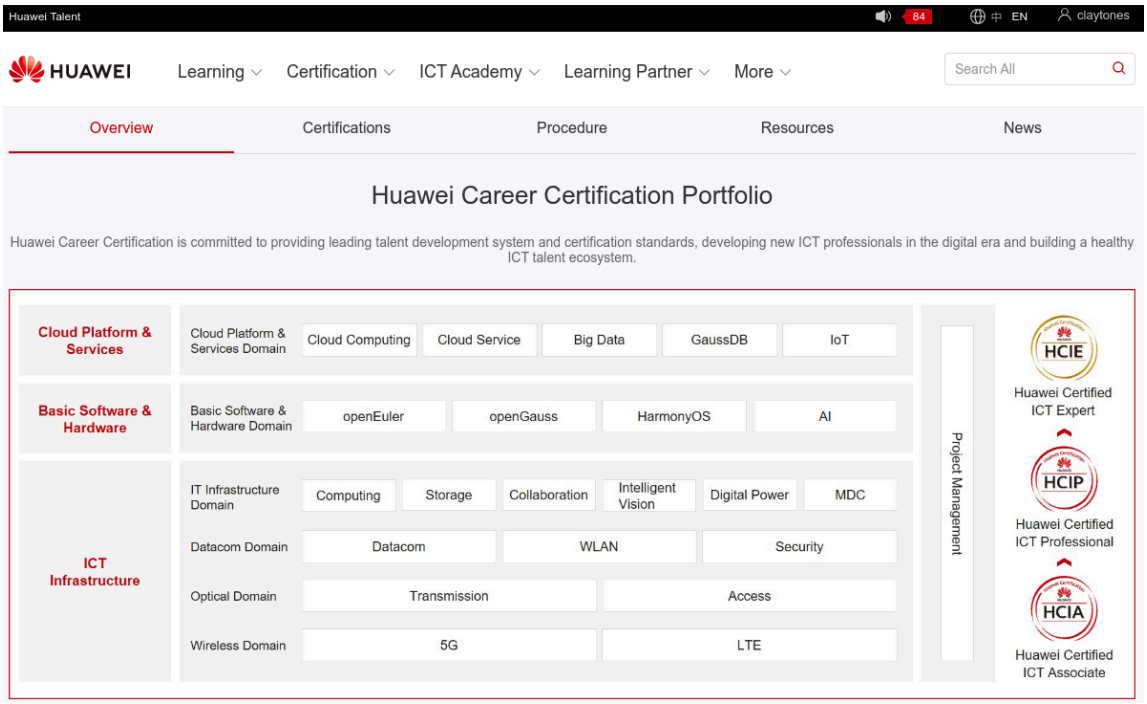


Figura 11.4: Visão geral do portfólio de certificações da Huawei

- ***Certificações:*** apresenta todo o portfólio de certificações profissionais oferecidas pela *Huawei*, que contempla todas as áreas de Tecnologia da Informação e Comunicação. Para informações detalhadas sobre a certificação desejada, basta clicar sobre o nome do exame de interesse;

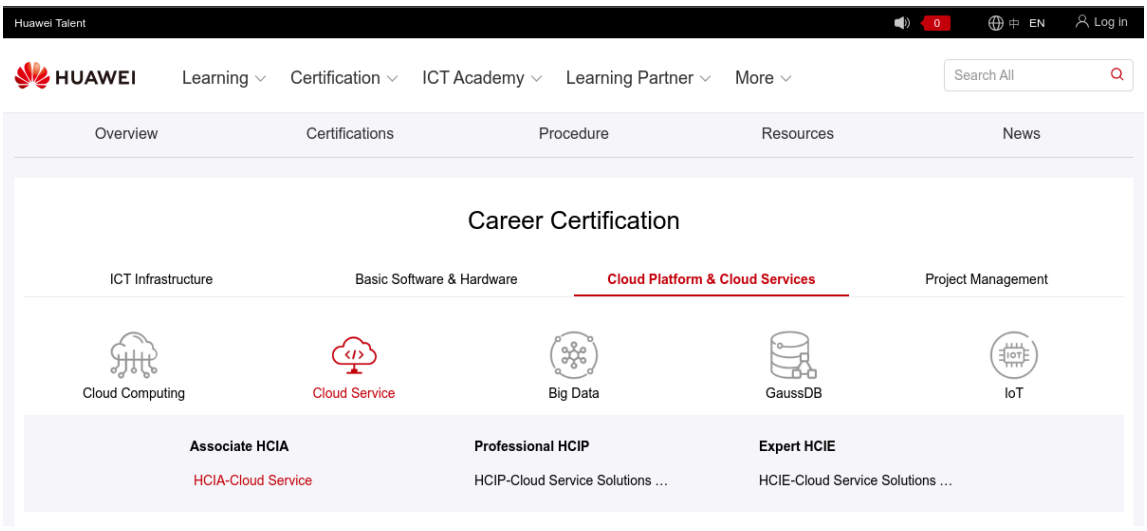


Figura 11.5: Certificações Cloud oferecidas pela Huawei

- ***Procedimento:*** detalha as etapas pertinentes aos processos de agendamento e efetiva realização dos exames de certificação. Também disponibiliza informações referentes à política de recertificação voltada a usuários já certificados que eventualmente precisem renovar certificações que estejam próximas do prazo de vencimento;

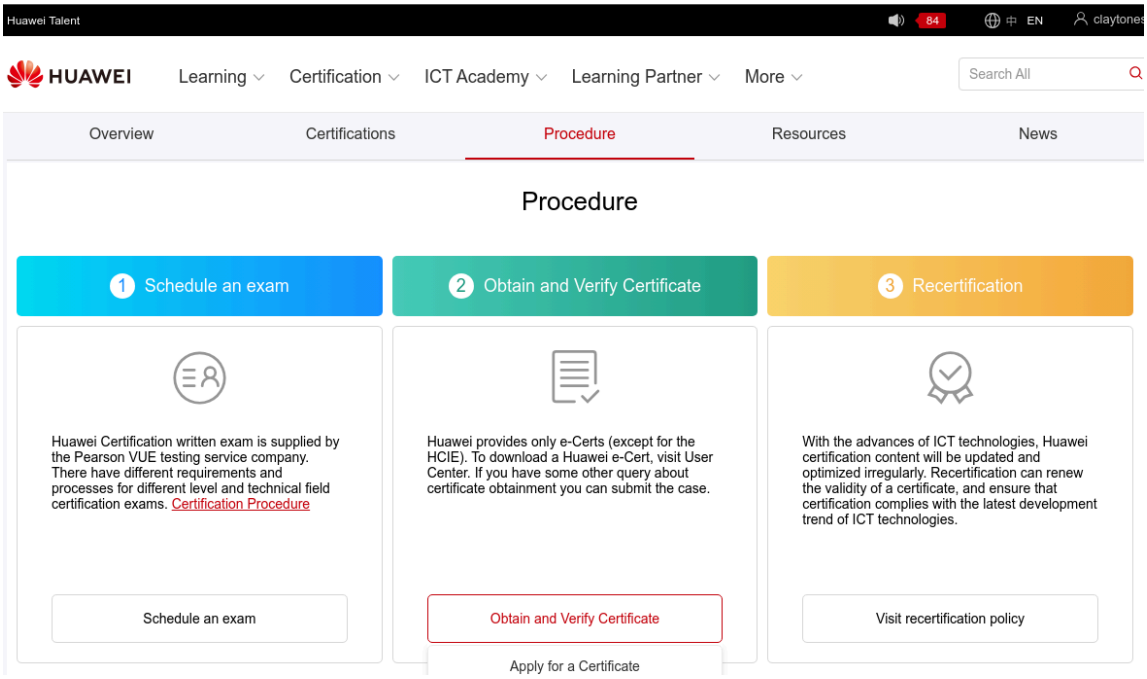


Figura 11.6: Visão geral do portfólio de certificações da Huawei

  - ***Recursos:*** nesta seção são disponibilizadas recomendações de recursos de aprendizagem existentes na plataforma, ofertadas aos estudantes. Entre os materiais disponíveis, destacam-se livros de referência, guias de estudo e treinamentos *online*;
- ***Notícias:*** traz informações atualizadas constantemente, relacionadas às certificações *Huawei*.

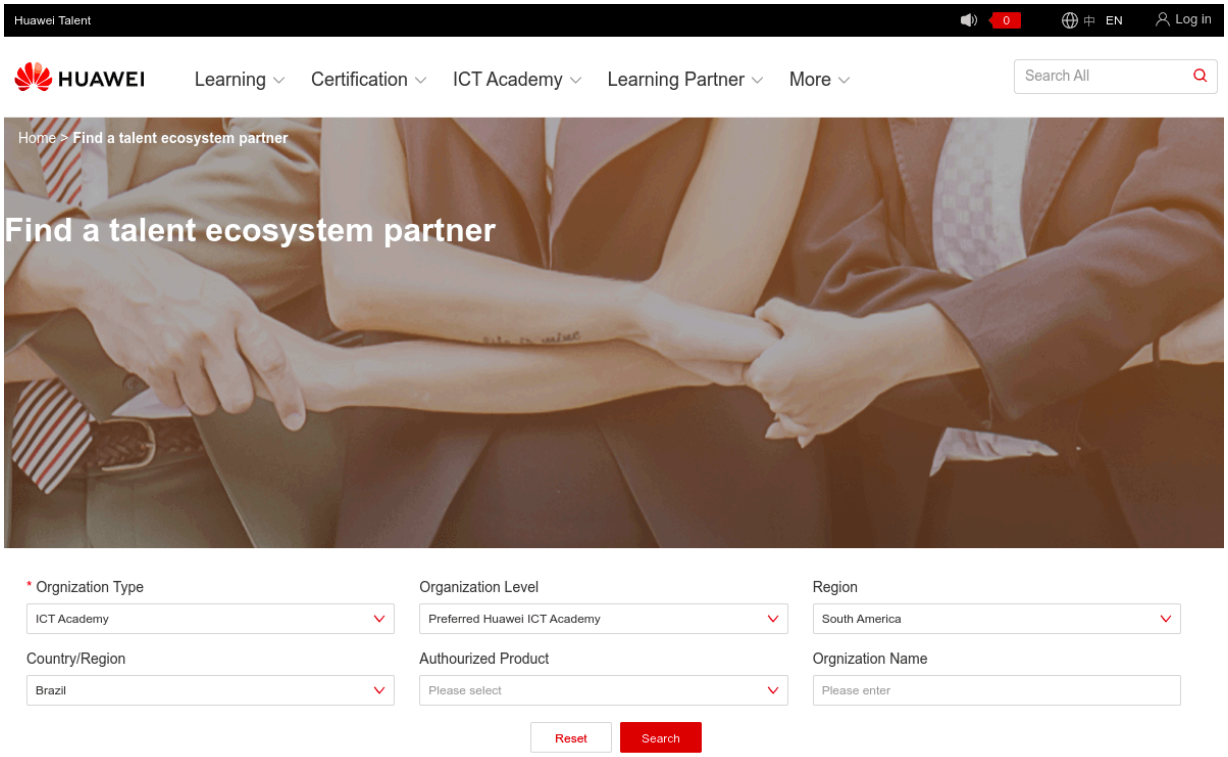
ICT Academy:

Para que o usuário visitante tenha acesso aos conteúdos disponibilizados na plataforma *Talent*, bem como a todos os demais serviços voltados aos clientes da *Huawei*, é necessária a criação de uma *Huawei ID*.

Uma das portas de entrada para este universo são as *ICT Academies*, nomenclatura que identifica as unidades participantes do programa acadêmico de mesmo nome criado pela *Huawei*, responsável por construir parcerias com instituições de ensino de todo o mundo. Por meio do programa, instrutores pertencentes às instituições de ensino parceiras ministram treinamentos oficiais, relacionados às principais tecnologias TIC da *Huawei*, para estudantes vinculados à sua Academia.

O programa tem como principal objetivo estimular estudantes a conquistarem certificações profissionais relacionadas a tecnologias relevantes, amplamente reconhecidas no mercado e, em especial, encontrar e desenvolver talentos, de modo a contribuir na construção de um ecossistema composto por cidadãos capazes de atuar na indústria e na sociedade, com habilidades diferenciadas.

Para verificar quais instituições de ensino fazem parte da iniciativa, basta consultar a ferramenta de busca específica para essa finalidade, disponível na opção *"Search ICT Academy"* do menu.



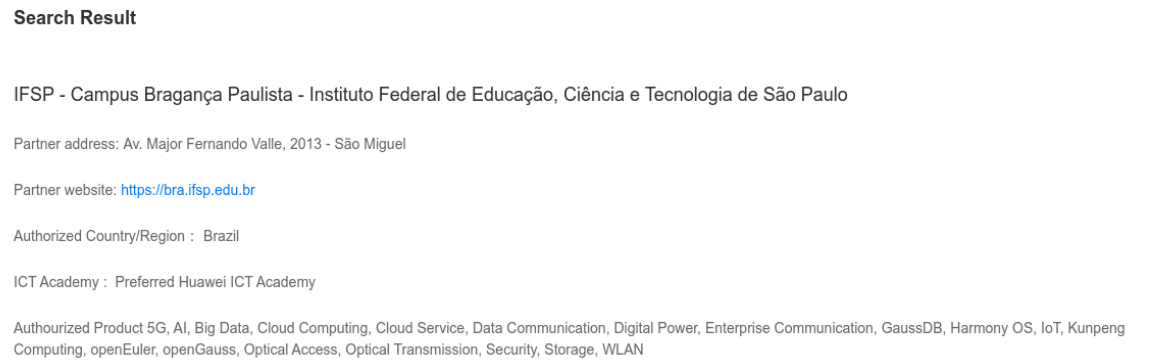


Figura 11.7: Huawei ICT Academy do campus Bragança Paulista do IFSP

Uma vez encontrada uma *ICT Academy* na região desejada, o interessado com vínculo ativo na instituição em questão pode iniciar o processo de inscrição como estudante da Academia. Um roteiro completo em vídeo, ilustrando todo o processo de inscrição, pode ser encontrado na *playlist* do livro, no canal *unixlabs* do *YouTube*.

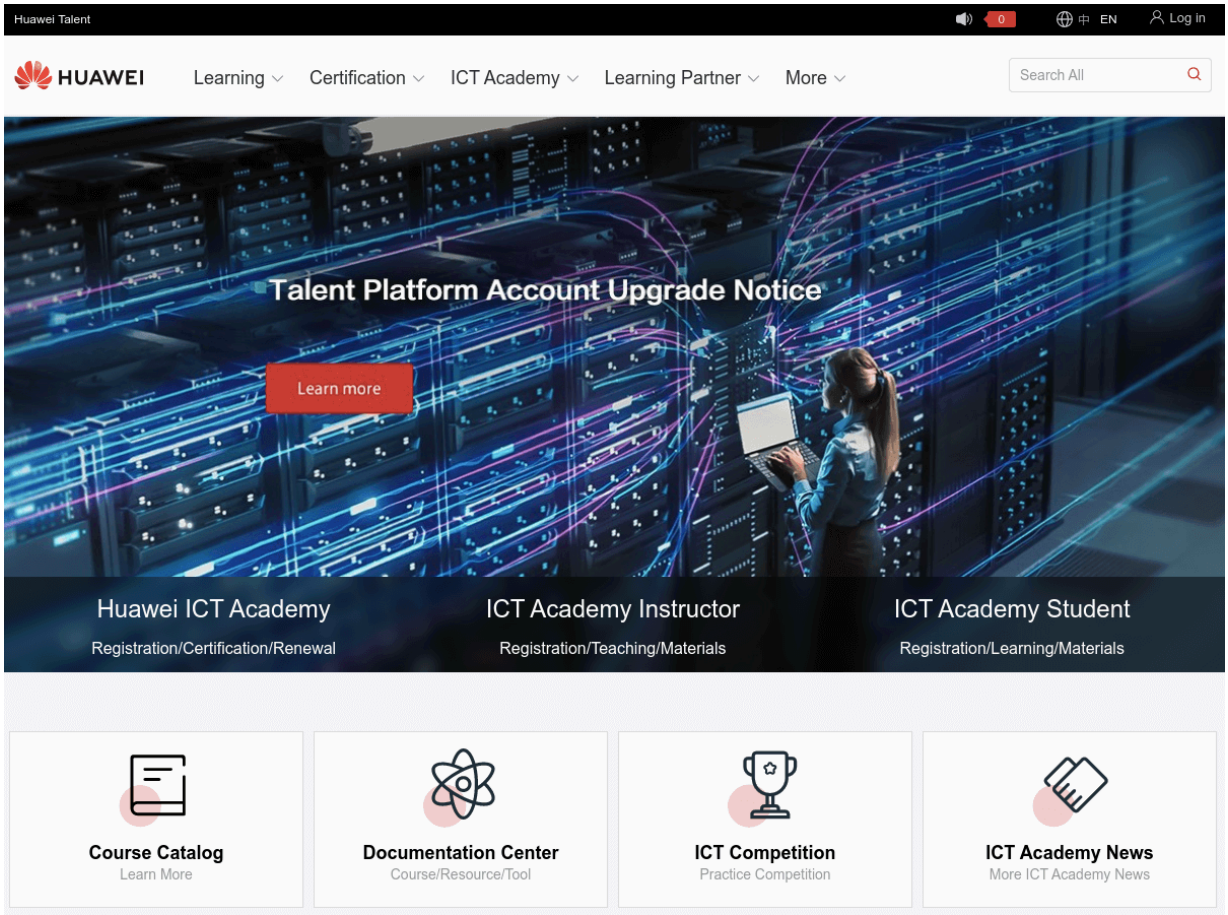


Figura 11.8: Cadastro de um estudante no programa Huawei ICT Academy

Para ter acesso aos conteúdos do programa, é necessário, primeiro, um cadastro na plataforma (passo 1), em seguida, o vínculo com uma academia ativa (passo 2) e, por fim, a matrícula no treinamento desejado (passo 3), utilizando um código de convite pertinente à turma desejada, devidamente fornecido pelo instrutor do treinamento.

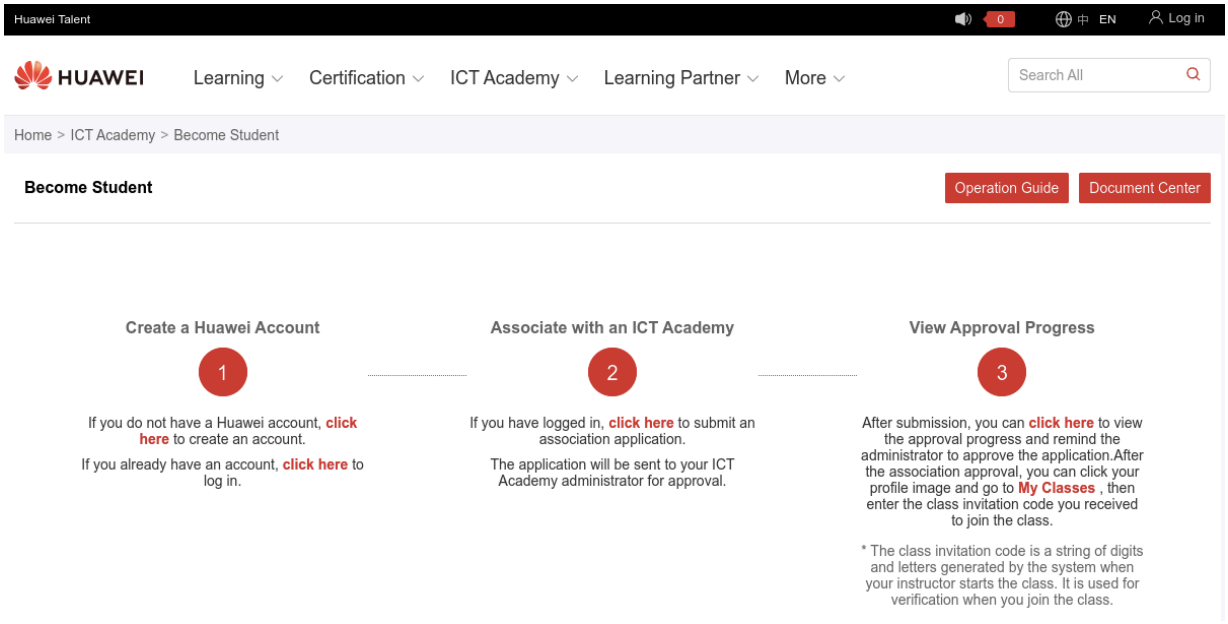


Figura 11.9: O processo de cadastro é dividido em três etapas

11.2 Plataforma KooLabs

Uma vez compreendidos os conceitos trabalhados em cada um dos capítulos da presente obra, é de fundamental importância complementar os conhecimentos adquiridos, associando a teoria utilizada em cada uma das principais tecnologias de nuvem a experimentos práticos baseados em cenários reais.

Para tanto, podemos utilizar a plataforma de laboratórios oficial da *Huawei Cloud*, denominada *KooLabs*, disponível para utilização gratuita no endereço https://lab.huaweicloud.com/intl/en-us/.

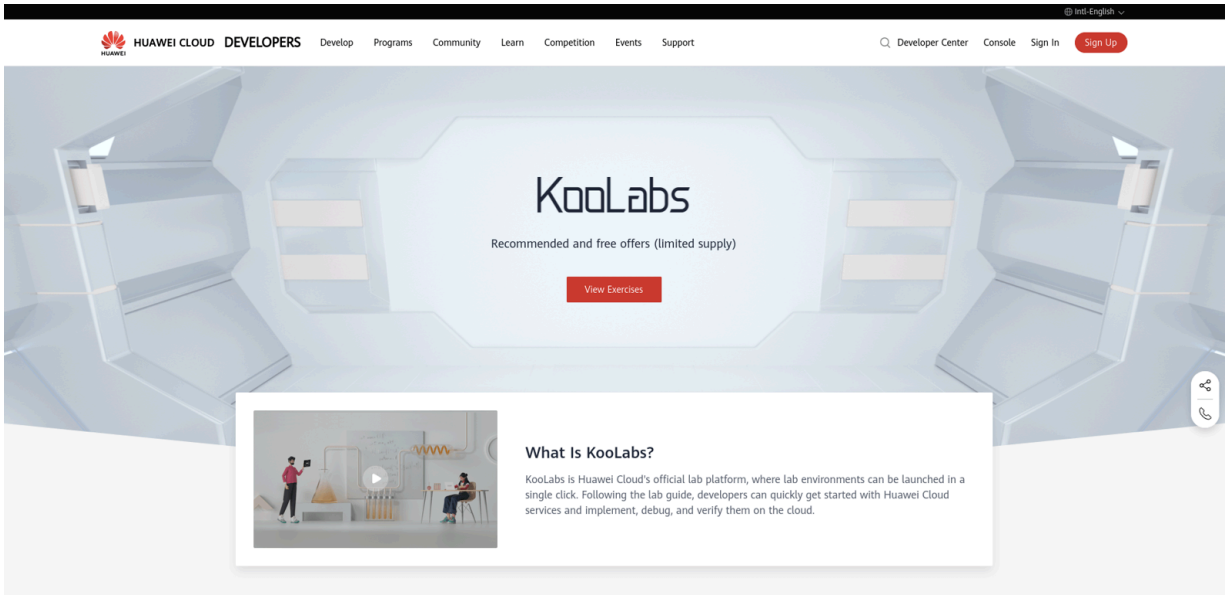


Figura 11.10: Plataforma de laboratórios práticos KooLabs, da Huawei Cloud

A plataforma oferece diversos exercícios práticos, relacionados aos principais e mais relevantes serviços disponibilizados pela *Huawei Cloud*, com diferentes níveis de complexidade.

No topo da página de visualização de exercícios, é possível filtrar as opções disponíveis em diferentes categorias — como *Databases, Compute, AI, Analytics, Internet of Things e Developer Services* — organizadas em cenários voltados para públicos de nível iniciante, intermediário ou avançado.

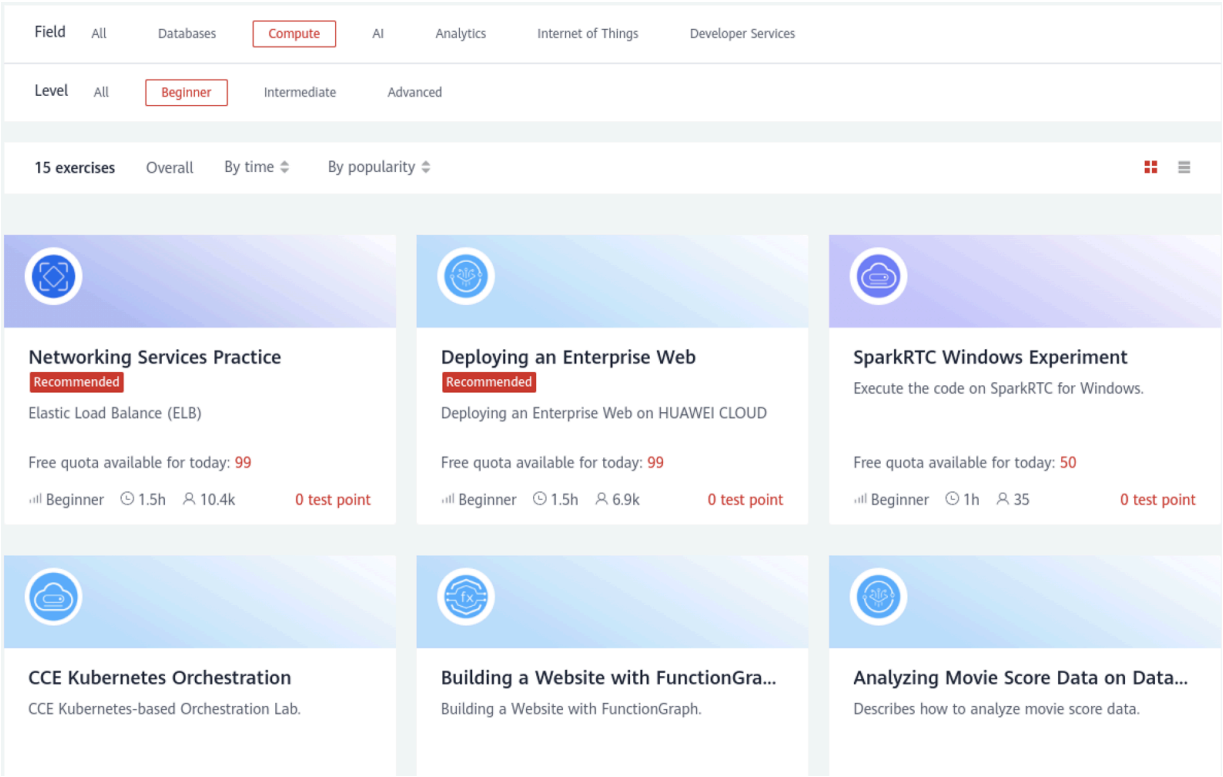


Figura 11.11: Áreas de atuação e níveis de complexidade disponíveis na plataforma de laboratórios práticos KooLabs, da Huawei Cloud

Uma vez definida a área de interesse e o nível de dificuldade da atividade que se deseja realizar, o usuário deve escolher uma das opções exibidas de acordo com sua preferência, e clicar sobre o título dela para que informações mais detalhadas possam ser disponibilizadas.

Dentre as informações exibidas, destacam-se os objetivos, requisitos e principais etapas da atividade, serviços *Huawei Cloud* utilizados e até mesmo cursos gratuitos de curta duração relacionados ao tema da prática, disponibilizados na plataforma.

Um outro fator importante que também deve ser observado é a duração estimada do laboratório, disponibilizada em minutos. Como se trata de uma plataforma gratuita, recomenda-se que as práticas sejam realizadas em momentos em que o estudante disponha minimamente do tempo informado como necessário para a realização da atividade. Tal prática possibilita que o exercício seja

feito com calma, de modo a consolidar os conhecimentos do praticante e evitar que créditos de nuvem sejam desperdiçados sem necessidade.

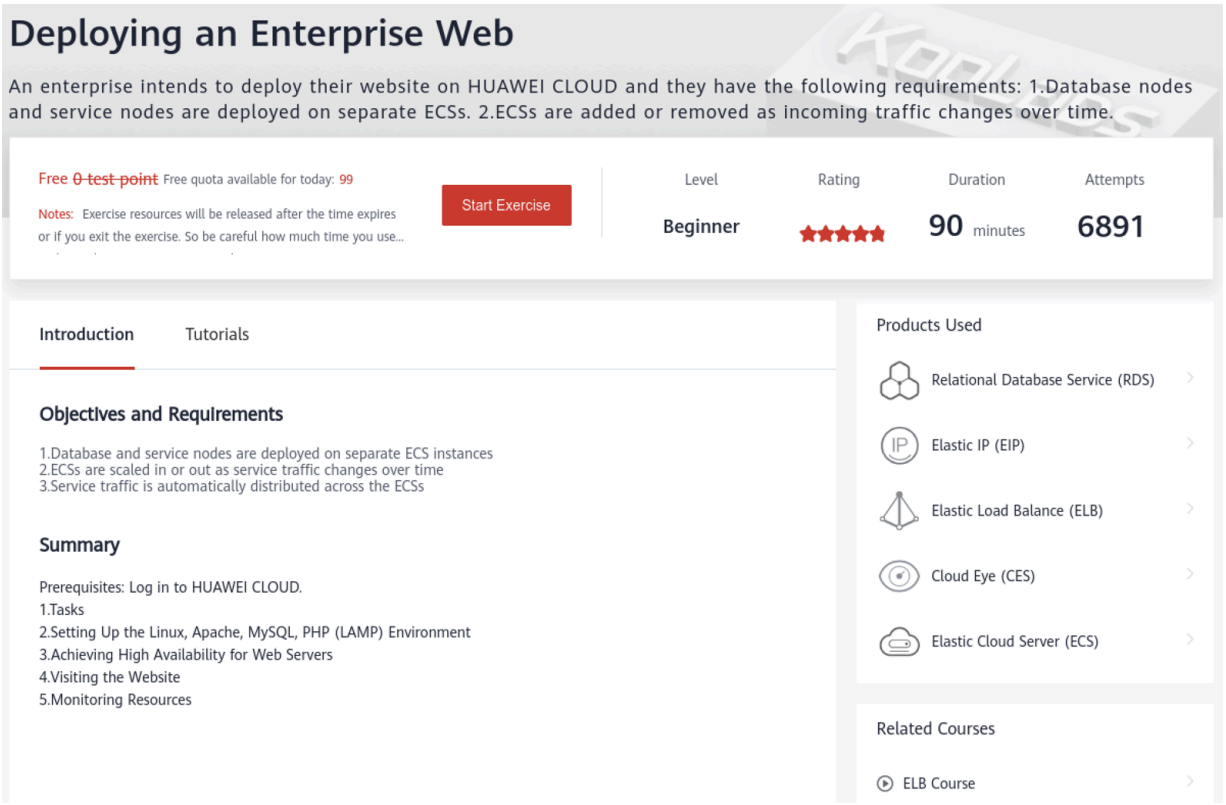


Figura 11.12: Informações detalhadas sobre a prática de laboratórios a ser realizada

Para iniciar a prática de laboratório, é necessário que o usuário se autentique na plataforma, utilizando suas credenciais *Huawei ID*.

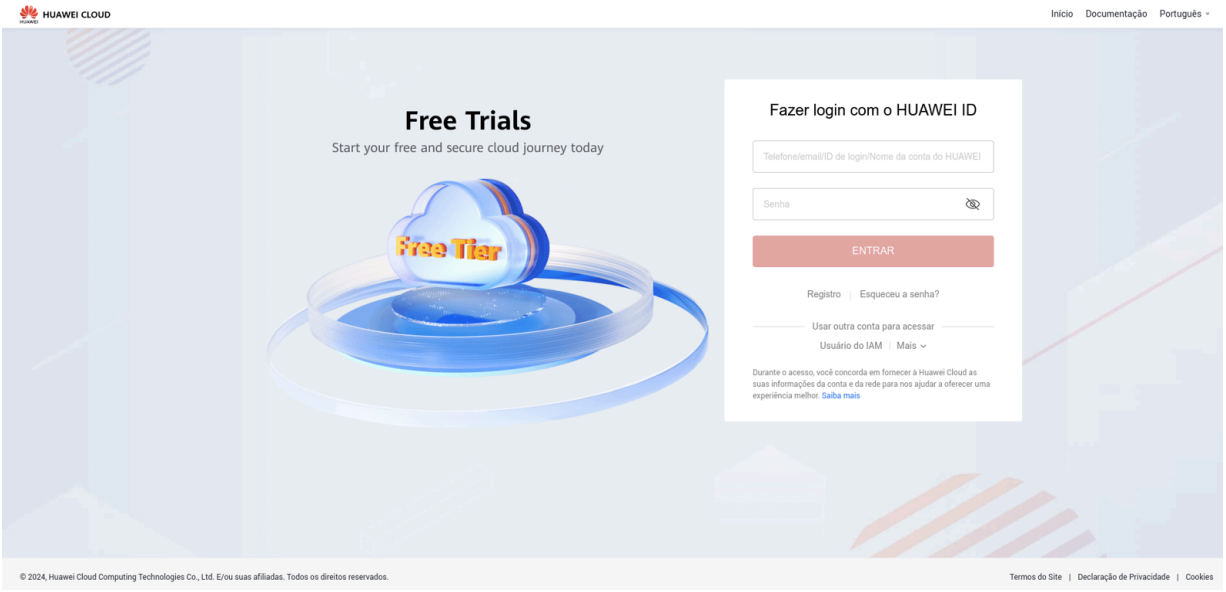


Figura 11.13: Autenticando-se na plataforma via Huawei ID

Uma vez autenticado, o usuário terá acesso à interface de trabalho pertinente ao laboratório escolhido, dividida em dois quadrantes distintos. O quadrante esquerdo contempla duas abas. A primeira, denominada *"Tutorials"*, traz as instruções necessárias para a realização do laboratório prático, no formato de um roteiro detalhado, passo a passo.

Figura 11.14: Tela inicial da atividade escolhida pelo usuário

A segunda aba, por sua vez, recebe o nome de *"Reports"* e detalha, dentre outras informações, o objetivo e o andamento individual de cada uma das atividades pertencente ao laboratório em execução, classificando-as como "completos" ou "não finalizadas".

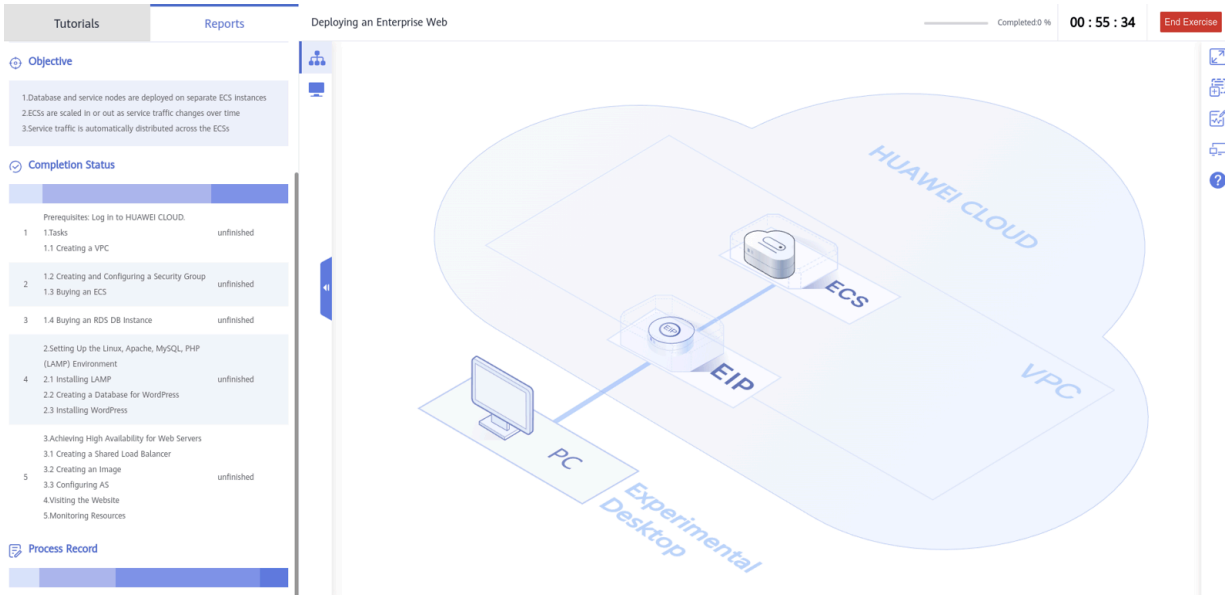


Figura 11.15: Relatórios de atividades realizadas na prática corrente

No quadrante direito, o estudante tem acesso a um espaço de trabalho mais amplo, que contempla, no topo direito, um *timer* que controla o tempo restante da atividade e um botão para encerramento do laboratório em execução. Logo abaixo, é possível ter acesso a duas interfaces distintas. A primeira, apresentada por padrão no momento em que o usuário acessa a interface, é representada por um ícone que alinha equipamentos de forma hierárquica e tem a função de exibir a topologia utilizada no laboratório. Para acessar a segunda interface, basta clicar na aba identificada pelo ícone de um pequeno monitor, responsável pelo acesso a uma conexão *Remote Desktop Protocol (RDP)* a uma área de trabalho onde, efetivamente, a atividade será realizada.

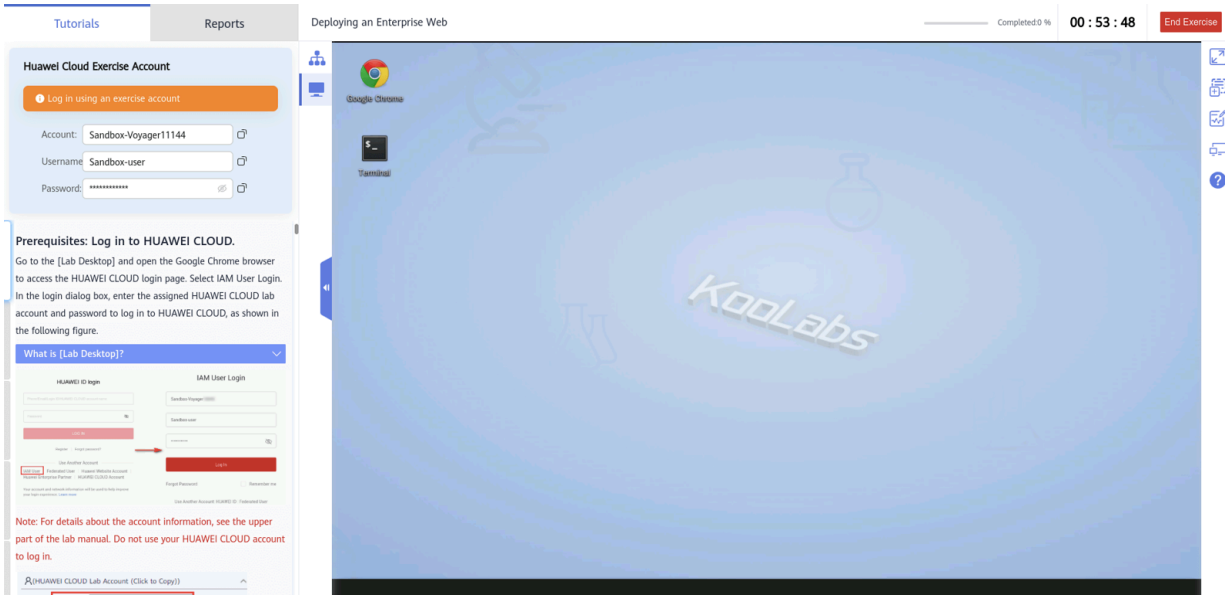


Figura 11.16: Acesso RDP remoto para realização da atividade

Para o laboratório em questão, será necessário clicar duas vezes no ícone do navegador *Google Chrome* apresentado na área de trabalho, para iniciar a atividade.

Em seguida, o usuário é levado novamente ao portal de autenticação da *Huawei Cloud*. Desta vez, as credenciais utilizadas devem ser correspondentes as de um usuário IAM válido, mais especificamente, as informadas no quadrante esquerdo, na aba *Tutorial*, constantes na seção *Huawei Cloud Exercise Account*. Não utilize suas credenciais pessoais nesse ambiente!

Vale lembrar que as credenciais em questão fazem parte de um *pool* de recursos dinâmicos, que são randômicos, válidos somente durante a vigência de tempo referente à sessão em execução e possuem permissionamento restrito, pertinente às atividades práticas associadas ao laboratório escolhido.

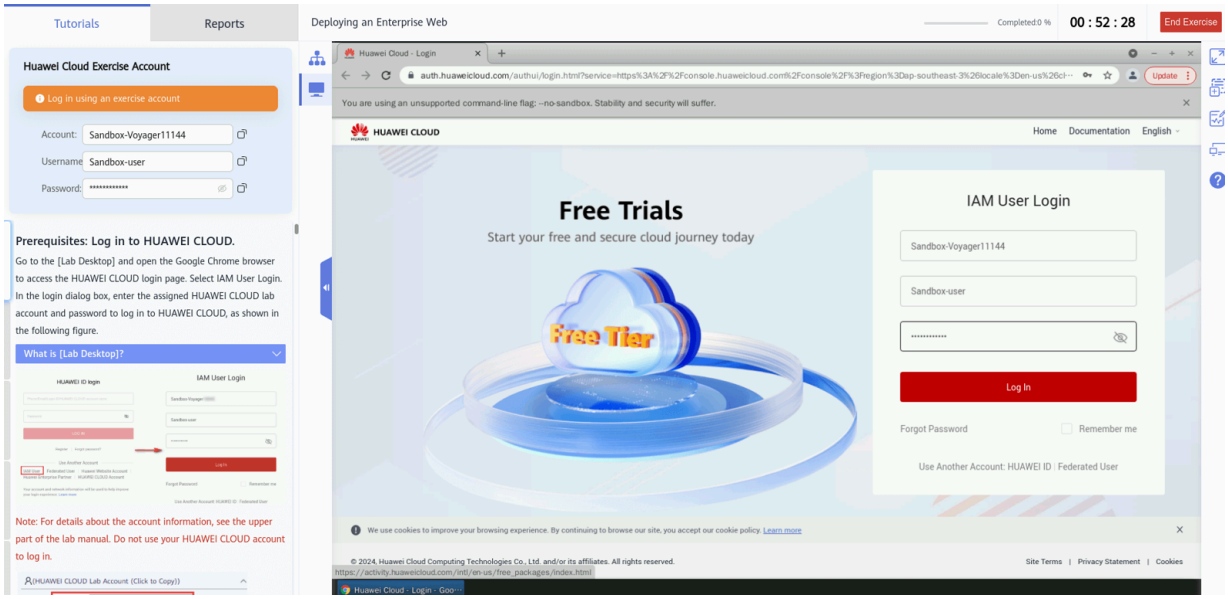


Figura 11.17: Autenticação na Huawei Cloud com as credenciais disponibilizadas na atividade

Por fim, o usuário tem acesso ao console de gerenciamento da *Huawei Cloud*, com um perfil *"sandbox-user"*, justamente voltado à finalidade de estudos e experimentação, com limitações aplicadas, portanto, a esse tipo de contexto.

Daqui em diante, o estudante deve seguir o roteiro, composto por instruções detalhadas, disponível na aba *Tutorials*, no quadrante esquerdo, e verificar o andamento das atividades por meio dos relatórios de acompanhamento disponibilizados na aba vizinha, denominada *Reports*.

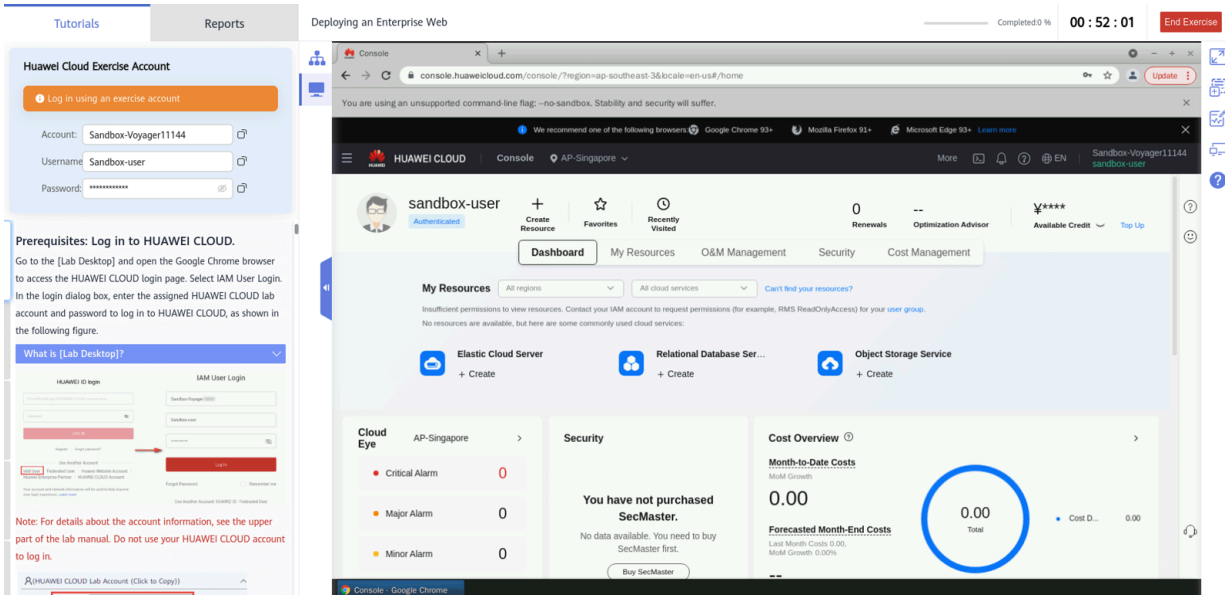


Figura 11.18: Início das atividades práticas de laboratório na Huawei Cloud

11.3 Considerações importantes

Com o objetivo de consolidar os conhecimentos adquiridos ao longo da leitura de cada capítulo e da realização das práticas sugeridas na plataforma de laboratórios oficial da *Huawei Cloud*, como cortesia, vamos disponibilizar um curso de extensão ofertado pela *Huawei ICT Academy* do campus Bragança Paulista do IFSP.

O treinamento em questão permite que o leitor ou leitora revise os conteúdos apresentados na presente obra e obtenha um certificado emitido pela plataforma *Huawei Talent*.

Para ser aprovada no "Exame Final", a pessoa candidata deve atingir uma pontuação mínima de 60% de acertos. A avaliação é composta por cerca de 65 questões de múltipla escolha sobre computação em nuvem, com um formato semelhante ao do exame de certificação correspondente.

Sendo assim, você tem a oportunidade de acessar conteúdos teóricos e práticos, testar seus conhecimentos e, eventualmente, considerar a possibilidade de realizar uma prova de certificação profissional, caso tenha interesse e disponibilidade para adquirir o *voucher* por meio de um centro de aplicação de exames devidamente credenciado pelo fabricante.

As informações sobre o treinamento estão disponíveis na seção "Recursos didáticos" deste capítulo. A cada semestre, uma nova turma será criada em substituição a do período anterior, e as informações correspondentes serão devidamente atualizadas. Esse procedimento garante que os conteúdos permaneçam atualizados a cada mudança de currículo.

As turmas continuarão sendo ofertadas durante o período em que o autor seguir atuando como instrutor credenciado e certificado no programa *Huawei ICT Academy*.

Questionário online

Figura 11.19: QR Code para acesso ao questionário online relacionado à temática do capítulo

Recursos didáticos

Figura 11.20: QR Code para acesso a materiais adicionais disponibilizados online relacionado à temática do capítulo